



UNIVERSITÀ DEGLI STUDI
DI TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE
ICT International Doctoral School

DOMAIN MODELING

THEORY AND PRACTICE

Subhashis Das

Advisor

Prof. Fausto Giunchiglia

Università degli Studi di Trento

Reviewer

Prof. Dave Robertson

The University of Edinburgh, UK

Reviewer

Asst. Prof. Biswanath Dutta

DRTC, Indian Statistical Institute, India

January 2018

Thesis committee

Assoc. Prof. Maurizio Marchese
The University of Trento, Italy

Thesis committee

Asst. Prof. Biswanath Dutta
DRTC, Indian Statistical Institute, India

Thesis committee

Research Associate Feroz Farazi
The University of Cambridge, UK

I hereby declare that I have created this work completely on my own and used no other sources or tools than the ones listed, and that I have marked any citations accordingly.

Trento, January 2018
Subhashis Das

Acknowledgements

I would not have been able to complete this thesis without help from many faculty members and students in the University of Trento. First and foremost I would like to thank Prof. Fausto Giunchiglia. I am so fortunate to have Prof. Fausto Giunchiglia as my advisor. Recalling the Ph.D. study in the past four years, I would say that I even could not complete my Ph.D. study without my advisor. He really did never give me up even during my most depressed time period for my Ph.D. study.

I would like to thanks my batch mate Sajjan Raj Ojha, Maria De Las Mercedes Huertas, Enrico Bignotti and seniors Dr. Vincenzo Maltese, Mattia Fumagalli, Amit kumar Sarangi, Abhinav Kumar Singh for their support during my work. Last but not the least I would like to thanks Prof. A.R.D. Prasad, Prof. Dave Robertson, Prof. Devika P. Madalli, Asst. Prof. Biswanath Dutta, Asst. Prof. Moumita Roy, Prof. Asis Ghosh, and my best friend Hasib-ur Rahaman Molla for their mental support during my hard time. Finally, my special thanks go to my extended network of family and friends including Mr. Sayon Roy, Preet Laal, Vaibhav Gharat, too countless to name, for their passionate and unconditional support. A person cannot achieve too much without help from others. You are all my friends who are my life-long fortune. Much love to you all. This research was funded by Erasmus Mundus India4EU II, SmartSociety, a project of the Seventh Framework Programme for Research of the European Community under grant agreement no. 600854 and QROWD, a project of the EU-Horizon2020 Programme for Research of the European Community under grant agreement no. 732194.

“No matter how many mistakes you make or how slow you progress, you are still way ahead of everyone who isn't trying.”
— Tony Robbins

Abstract

Everyday huge amount of data is being captured and stored. This can either be due to several social initiatives, technological advancement or by smart devices. This involves the release of data which differs in format, language, schema and standards from various types of user communities and organizations. The main challenge in this scenario lies in the integration of such diverse data and on the generator of knowledge from the existing sources. Various methodology for data modeling has been proposed by different research groups, under different approaches and based on the scenarios of the different domain of application. However, a few methodology elaborates the proceeding steps. As a result, there is lack of clarification how to handle different issues which occurs in the different phases of domain modeling. The aim of this research is to presents a scalable, interoperable, effective framework and a methodology for data modeling. The backbone of the framework is composed of a two-layer, schema and language, to tackle diversity. An entity-centric approach has been followed as a main notion of the methodology. A few aspects which have especially been emphasized are: modeling a flexible data integration schema, dealing with the messy data source, alignment with an upper ontology and implementation. We evaluated our methodology from the user perspective to check its practicability.

Keywords

[Data management, Diversity, Domain, Data integration, Geospatial, Healthcare, Methodology, Ontology, Transportation]

Publications

- Ojha S. R., **Das, S.**, & Karanjit, S. (2018). A Process Ontology for a confectionery service robot. *International Journal of Semantic Computing*, 12(01).
- **Das, S.**, Ojha, S. R., & Giunchiglia, F. (2017, January). ATOM: Ontology Aware Transportation Model. In *Semantic Computing (ICSC), 2017 IEEE 11th International Conference on* (pp. 272-273). IEEE.
- Giunchiglia, F., Ojha, S. R., & **Das, S.** (2017, January). SemUI: A Knowledge Driven Visualization of Diversified Data. In *Semantic Computing (ICSC), 2017 IEEE 11th International Conference on* (pp. 234-241). IEEE.
- Karanjit, S., Ojha, S. R., & **Das, S.** (2017, January). Process Ontology for Confectionery SweetBot. In *Semantic Computing (ICSC), 2017 IEEE 11th International Conference on* (pp. 378-383). IEEE.
- **Das, S.**, & Giunchiglia, F. (2016, October). GeoeTypes: Harmonizing Diversity in Geospatial Data (Short Paper). In *OTM Confederated International Conferences "On the Move to Meaningful Internet Systems"* (pp. 643-653). Springer International Publishing.
- **Das, S.**, & Roy, S. (2016). Faceted Ontological Model for Brain Tumour Study. *Knowledge Organization*, 43(1).
- Roy, S., & **Das, S.** (2015, January). Spatial data infrastructures: Its metadata and analysis. In *Emerging Trends and Technologies in Libraries and Information Services (ETTLIS), 2015 4th International Symposium on* (pp. 43-51). IEEE.

Contents

1	Introduction	1
1.1	The Context	2
1.2	The Problem	3
1.3	The Solution	5
1.4	Structure of the Thesis	6
2	Methodology for Domain Modeling	9
2.1	Domain knowledge	11
2.2	Reference Scenarios	12
2.3	Generalized Queries	13
2.3.1	Identify the Concepts	13
2.3.2	Query Generation Methodology	14
2.4	State of the Art	14
2.4.1	Standards	14
2.4.2	Resources	15
2.4.3	Applications	16
2.5	Informal Modeling Phase	16
2.5.1	Pilot Reference Datasets	17
2.5.2	ExER Model	17
2.6	eTypes Model	17
2.7	Schema Level	20
2.8	Language Level	24

2.9	Evaluation	25
2.9.1	Model Evaluation	25
2.9.2	Terminology Evaluation	28
2.10	Case Studies	28
2.11	Related Work	28
2.11.1	Ontology Development Methodology	29
2.11.2	Ontology Evaluation Technique	33
3	Geospatial Domain	37
3.1	Domain Description	37
3.1.1	Basic Definition	39
3.1.2	Purpose and objectives	42
3.1.3	Problems	42
3.2	Reference Scenarios	43
3.3	Generalized Queries	44
3.3.1	Identify the Domain Concepts	45
3.3.2	Query Collection	50
3.4	State of the Art	51
3.4.1	Standards	51
3.4.2	Resources	52
3.4.3	Applications	55
3.5	Informal Modeling Phase	58
3.5.1	Pilot Reference Datasets	58
3.5.2	ExER Model	58
3.6	Geo eTypes	59
3.7	Schema Level	63
3.7.1	Defined Classes	63
3.7.2	Alignment with Upper Ontology	65
3.7.3	Defined Attributes	66

3.8	Language Level	68
3.9	Evaluation	69
4	Smart Transportation Domain	73
4.1	Domain Description	73
4.1.1	Problems	75
4.1.2	Purpose and Objectives	76
4.2	Reference Scenarios	76
4.3	Generalized Queries	78
4.3.1	Identify the Domain Concepts	78
4.3.2	Query Collection	80
4.4	State of the Art	80
4.5	Informal Modeling Phase	81
4.5.1	Pilot Reference Datasets	81
4.5.2	ExER Model	81
4.6	Transportation eTypes	82
4.7	Schema Level	83
4.8	Language Level	87
4.9	Evaluation	87
5	Healthcare Domain	93
5.1	Domain Description	93
5.2	Reference Scenarios	95
5.3	Generalized Queries	95
5.3.1	Identify the Domain Concepts	95
5.3.2	Query Collection	97
5.4	State of the Art	97
5.4.1	Standards	97
5.4.2	Applications	102
5.5	Informal Modeling Phase	106

5.5.1	Pilot Reference Datasets	106
5.5.2	ExER Model	107
5.6	Health eTypes	108
5.7	Schema Level	112
5.7.1	Defined Classes	112
5.7.2	Alignment with top level ontology	115
5.8	Language Level	116
5.9	Evaluation	116
5.10	Issues	121
6	Evaluation	123
7	Implementation	127
8	Conclusion and Future Work	131
8.1	observation	132
8.1.1	Stakeholder Analysis	132
8.1.2	Workflow Integration	133
8.2	Lesson Learned	134
8.2.1	Extending the WordNet with Domain Knowledge .	134
8.2.2	Deciding the Import Order	135
8.2.3	Attaching Domain sub-tree	135
8.2.4	Specifying Concept Metadata	135
8.3	Future Work	136
	Bibliography	137
A	Authority File	147
A.1	Tools	147
A.2	Geo Authority File	148
A.3	eTypes Specification	149

List of Tables

3.1	Pilot Geo data sources	58
3.2	Location eType	63
3.3	Administrative district eType	63
3.4	Building eType	64
3.5	Spatial Attribute	67
3.6	Temporal Attribute	68
4.1	Trip eType	83
4.2	Parking eType	84
5.1	Healthcare Dataset 1	107
5.2	Healthcare Dataset2	107
5.3	Prescription eType	110
5.4	Patient eType	111
5.5	Visit eType	111
5.6	Person eType	112
A.1	OpenStreetMap	149
A.2	GeoNames	150
A.3	Specimen eType	150
A.4	Death eType	151
A.5	Visit eType	151
A.6	Provider eType	151
A.7	Disease eType	152

List of Figures

1.1	Different semantic impediments related with data in any Open Data ecosystem which creates an obstruction for easy consumption of the data.	5
1.2	Structure of the thesis	7
2.1	Procedure of Domain Medeling	10
2.2	Domain Knowledge	11
2.3	eType Metamodel	20
2.4	Relation between schema level and language level	25
3.1	Geospatial Domain	38
3.2	ExER Model for Geospatial	59
3.3	Simple Overlay	60
3.4	Many-to-many linking	60
3.5	Referencing a common base	61
3.6	Main classes of Geo eTypes	65
3.7	Class Relationship Geospatial	66
3.8	Example of Facets	69
3.9	Statistic	70
3.10	Comparison of service quality and cuisine quality of restaurant in Trento	72
4.1	Transportation Domain	74
4.2	ExER Model for Transportation	82

4.3	Transportation class Hierarchy	85
4.4	Class Relationship-Transportation	86
4.5	Statistics Transport	88
4.6	SPARQL query for wheelchair access	89
4.7	SPARQL query for bikes allowed	90
4.8	SPARQL query for near by services	90
4.9	Transportation Knowledge Graph	91
5.1	Healthcare Domain	94
5.2	RxClass Browser	104
5.3	Human API	105
5.4	DrugBank Browser	106
5.5	ExER Model for Healthcare	108
5.6	Class Hierarchy of Healthcare	113
5.7	Class Relationship-Healthcare	115
5.8	comparison chart showing prescribe quantity of Drug X in respect with their age	118
5.9	Patient personal information along with the prescribe drugs	119
5.10	Partial view of Healthcare Knowledge Graph	121
5.11	Comparison of ICD-9 and ICD-10	122
6.1	User Experience dimensions. Adapted from [Laugwitz et al., 2008].	124
6.2	Result of Methodology Evaluation	125
7.1	Layered Architecture	128
7.2	SemUI	129
7.3	SHIB User Interface	130

Chapter 1

Introduction

“Coming together is a beginning; keeping together is progress; working together is success.” —*Henry Ford*

A domain represents concepts which belong to part of the world. Particular meanings of terms applied to that domain are provided by domain-specific resources. For example, the word school has many different meanings. A resource about the domain of education would model the “school building” meaning of the word, while a resource about the domain of fishing would model the “group of fish” meaning. The explicit meaning of a word can be capture using a domain-specific ontology.

Since domain ontologies represent concepts in very specific and often eclectic ways, they are often incompatible. As systems that rely on domain ontologies expand, they often need to merge domain ontologies into a more general representation. This presents a challenge to the ontology designer. Different ontologies in the same domain arise due to difference in domain-language, intended usage of domain, and domain perception.

Until now, several number of methodology has been proposed by different research groups, using different approaches and based on the scenarios

of different domain of application. However a few methodology have been suggested how to proceed, describing in the details steps. As a result, there is lack of clarification how to handle different issues which occurs in the different phases of the domain development.

1.1 The Context

Domain-specific standards and literature uses domain-specific schema and terminology which might different from general purpose resources. On the other hand general purpose resources including schema, control vocabulary or technical specifications are widely used in many domain application but it might not be sufficient for very domain-specific applications.

Purpose of our work is to bridge the gap among formal specification, practical uses of ontology and data availability by using international guidelines and a robust methodology which is missing most of the state of the art application. This work has the following objectives

- Explicitly clarify the scope by defining the core terms in the domain
- Identify basic entity types as well as domain-specific entity types
- Identify the attributes and relations, which are essential to describe them and enable domain application.
- Defining a model that specifies the terminology, which is needed for such entity types and their attributes.
- Test the model on various datasets (most likely from Trentino and Scotland), e.g. for integration purposes.

1.2 The Problem

We investigated various datasets in an effort to better understand different semantic and user related impediments existing in the domain-specific data which are mainly available as a Open Data or private data. The investigation involved consulting various data portals^{1,2}, listing the datasets that we intended to use, understanding their formats, finding terms used in the datasets to denote the different kind of features that matches our need, and finally integrating them. Our research questions were: (1) what kind of challenges would be encountered (specifically related to data and with the end users) and (2) what would be the requirements to build a common generic model for domain-specific data. Figure 1.1 shows different semantic impediments that exist in available Data. Issues related to domain-specific data is mainly characterized into the following categories:

- *Multidimensional Data.* Data are complex and related with a number of properties such as spatial, temporal or inherence. Dataset publishers can define their own classification and categories. For example, in Figure 1.1 there are many datasets belonging to one category. Merging these datasets for generating a unified model can be problematic.
- *Multilingualism.* Dealing with a multilingual dataset is in itself a big challenge. Global datasets such as OpenStreetMap datasets, United Nation's datasets are usually available in English but country-specific datasets are localized in its native language. For example, in Figure 1.1, the information about river Danube exists in different languages such as German, Hungarian or Croatian.
- *Diversity in the nomenclature.* Datasets published from different government and private organizations often use various labels or tags for

¹<http://opendata.arcgis.com>

²<https://www.europeandataportal.eu>

the same concept. This variation creates terminological diversity. For example, in Figure 1.1 *cap* or *zip* have similar meaning. Moreover, existing information systems use their own schema ignoring fundamental ontological principles. As an instance, many data models use different notations to represent the same attribute. In Figure 1.1 sometimes the data providers use X for denoting latitude and Y for longitude while in some other cases they use geo-coordinates to denote latitude and longitude together. This kind of name/tag variation is very difficult for a machine to understand without manual intervention.

- *Diversity in meaning.* Natural language description helps people to understand the intended meaning of the concept used in the schema. However, these natural language descriptions are usually ignored in information systems. For example in the case of the OpenStreetMap (OSM), no specific definition is available for many tags. Moreover, within the organization (e.g. finance department and administrative department) the same term is used to denote different concepts. For example, the meaning of the term “capital” can be “*assets available for use in the production of further assets*” or “*a seat of government*”; the meaning of the term “cold” can be “*a mild viral infection involving the nose and respiratory passages (but not the lungs)*” or “*the sensation produced by low temperatures*” .
- *Variation in data formats.* Data exists in multiple formats in an Open Data portal as shown in Figure 1.1. The end user has to know how to manipulate these files with different formats. It makes the task cumbersome to the end user.
- *Complexity in data correlation process.* Finding relationships within the data is a difficult task for the end users. For example, as in Figure 1.1, information about a hotel’s location is in a land register file

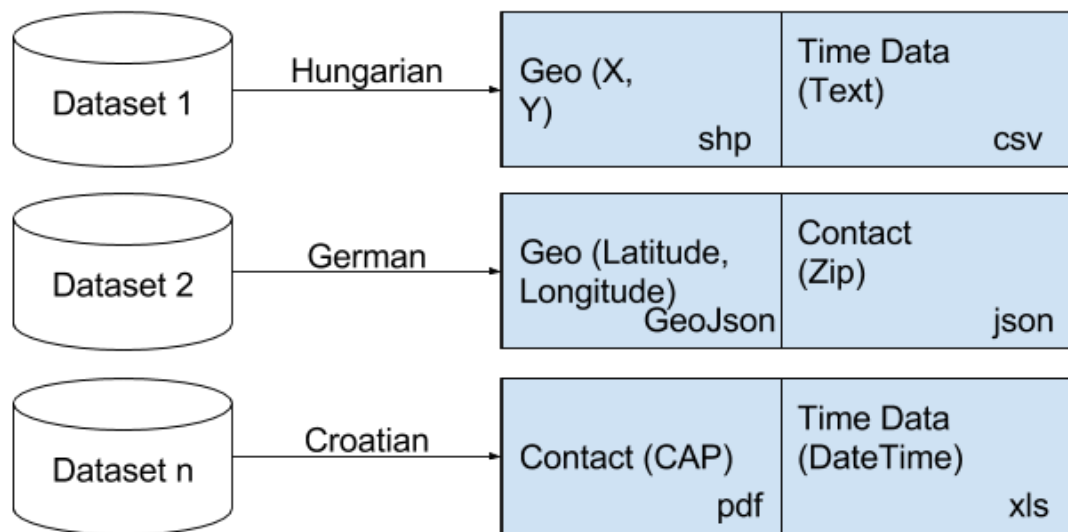


Figure 1.1: Different semantic impediments related with data in any Open Data ecosystem which creates an obstruction for easy consumption of the data.

whereas contact information is available in the tourist information file. The end user may not be able to comprehend the existing relationship between those two files. Furthermore, they might be stored in different categories (domains) complicating the task of finding the datasets.

1.3 The Solution

The contribution of this thesis is threefold: we (i) introduce a framework for capturing domain knowledge, (ii) we organize existing work in domain development methodology within this framework and fill missing spots, and finally (iii) we implement the theoretical results in practical application systems to validate our approach.

- A framework for capturing domain knowledge: Framework and content in domain modeling research has been fairly diverse. Section 2.11 contain a survey of literature, and consolidated the steps used. It identifies and defines a concise set of steps and development aspect

that can be evaluated (see Chapter 6).

- A model for domain knowledge: We clearly distinguished conceptual layer of the model (i.e. Schema Level) from natural language layer (i.e. Language Level). This give us maximum flexibility to handle diversity separately in schema and language without interfering both layer (see Section 2.6 to 2.8).
- Implementation: In order to check the theoretical framework of the domain development. We implemented it into three domain-specific application. A three-layer system architecture has been used to accommodate the model. We also test our proposed methodology in a class room environment to develop various domain as part of the student project. In the end we evaluated quality of our methodology from user perspective and result of preliminary two experiment was very promising.

1.4 Structure of the Thesis

The thesis report is structured in seven chapters Chapters 2 introduce the theoretical foundation and the methodology in which the work presented in this thesis has been carried out. Chapters 3 to 5 report the three main case studies around which this thesis is validated. Chapter 6 we provided quantitative evaluation of our proposed methodology based on User Experience (UX) dimension. Chapter 7 briefly describe the tool where we implemented our framework. Chapter 8 concludes with a discussion and future directions of the ongoing work.

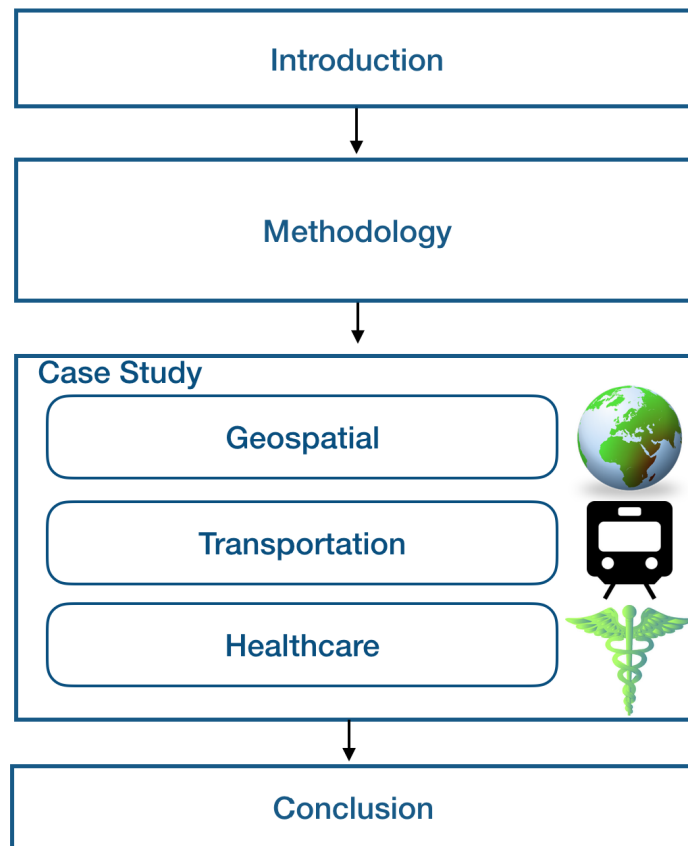


Figure 1.2: Structure of the thesis

Chapter 2

Methodology for Domain Modeling

“I am not here to
speak the Truth.
I am here just to give you a
method to perceive it.”
—*Sadhguru, Of Mystics &
Mistakes*

Until now, several number of methodology has been proposed by different research groups, under different approaches and based on the scenarios of different domain of application. However a few methodology have been suggested how to proceed, describing in the details steps. As a result, there is lack of clarification how to handle different issues which occurs in the different phases of domain development.

In this chapter, We introduce a framework for domain modeling. The rest of this thesis will be built on the framework described in this chapter. First, we give an overview of the domain in Section 2.1, introducing the relevant terms and their connections. Section 2.2 describes a reference application scenario for domain development. Based on that scenario, we define steps which we have to follow to build the model and language of the domain. We then describe the related work in Section 2.7. State of the art for the specific domain mentioned in the corresponding domain chapter.

Our proposed methodology is inspired by the Gruninger and Fox methodology [Gruninger, 1995], METHONTOLOGY [Fernández-López et al., 1997], and DERA methodology [Giunchiglia et al., 2014]. Main purpose is to reuse the existing methodology as much as possible and then adapt or adopt and extend when necessary.

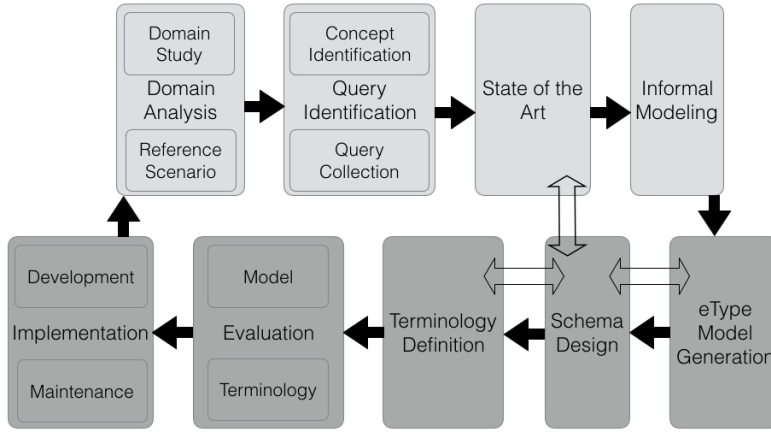


Figure 2.1: Procedure of Domain Modeling

Our methodology (see Figure 2.1) can be divided into three broad category. They are:

- **Informal:** In this phase we mainly deals with the natural language terms collected during generalized queries based on the reference scenarios and then look at the current state of the art to build preliminary model of the domain.
- **Formal:** In this phase main task is to formalize the model using formal language, define rules and constrain. Develop data model, acquired terminology.
- **Implementation:** Final phase is to implement the developed model in a real system. This enables us to very usability, accessibility, and understandability.

2.1 Domain knowledge

We can find several definition of domain in the state of the literatures among them we listed few of them which we used before defining our own definition of the domain. Bentivogli et. al[Bentivogli et al., 2004] defined a domain as “an area of knowledge which is somehow recognized as unitary. A domain can be characterized by the name of a discipline where a certain knowledge area is developed (e.g. chemistry) or by the specific object of the knowledge area (e.g. food). Although objects of knowledge and disciplines that study them are clearly related, the relation between these two points of view on domains is sometimes blurred and may be a source of uncertainty on their exact definition”.

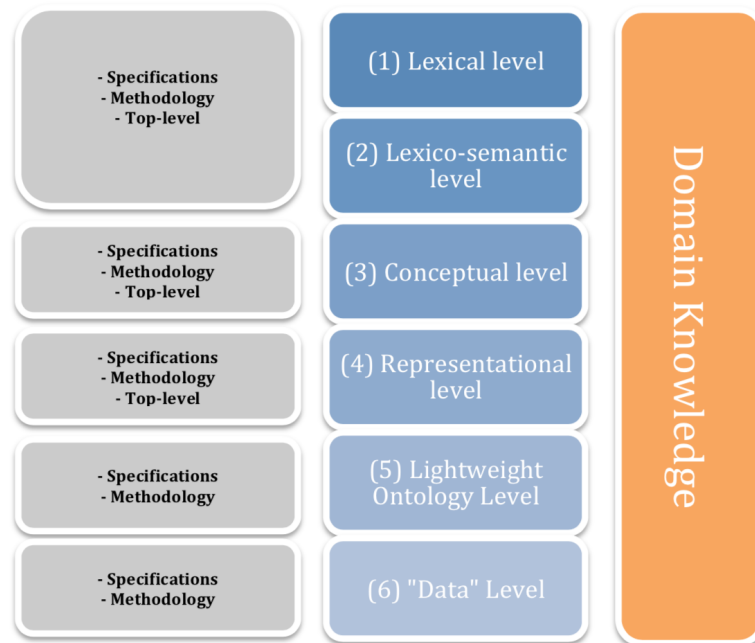


Figure 2.2: Domain Knowledge

By domain knowledge we understand formally or semi-formally represented knowledge resources about a specific field of study originating from authoritative sources. Each knowledge instance, i.e. any Knowledge Core

along the six levels, is what we call a domain. Figure 2.2 showing how domain knowledge can be store in the different level. The first two level, Lexical and Lexico-semantic level are associated with linguistic knowledge. The level-1 captures domain knowledge in from word or small sentence express in a natural language (i.e. English, Italian) label and the level-2 stores relation among those word using lexico-semantic relation (i.e. hyponym/hypernym). Middle two level, Conceptual and Representational level are language independent layer which represent knowledge in one or many abstract way (i.e. UML, ER or graph). Level five could be seen as a layer for storing lightweight ontology if available. Bottom layer is for row data from where we get true reflection of the domain. It is very necessary to understand different level of the domain to formulate the model and implement the domain application.

2.2 Reference Scenarios

The development of model and its terminology should be motivated by scenario that arise in the applications. In particular, such application scenarios may be presented by government authorities or industrial partners as problems which they face in their organization. The reference scenarios often have the from of story problems or examples which are not sufficiently tackle by existing domain models. A reference scenario also provides a set of guideline for possible solution to the scenario problems. These steps provide a rough idea of the intended model for the objects and relations that will later be included in the model.

2.3 Generalized Queries

Given the reference scenario, a set of queries will generate which place demands on an underlying model. We can consider these queries to be necessity that are in the form of questions that the model must be able to answer. These are the informal questions, later we need to build model by formalizing those concepts arise from the set of questions been collected from the query collections techniques.

By analyzing and specifying the set of competency questions and the application scenario, we make an overall structure for the new or extended model. This also set up and fixed the scope of our model; the analysis must be determined whether there is a need for the new model from scratch or extension of the existing model will be sufficient. Preferably, the generalized queries should be defined in a categorized manner, with higher level queries requiring the solution of more specific queries.

2.3.1 Identify the Concepts

By systematic analysis of queries, we can be able to understand main basic terminology used in that particular domain. This is an informal process of the methodology where we usually process natural language questions.

In order to get an idea about all terminology in a given domain, the first step in the process was the selection of one or more suitable sources of domain-specific terms. In principle, there are various ways to collect such terms. For example, this can be done by extracting them from texts on the domain-specific literature, by analyzing the millions of user queries stored in the query logs of existing search engines, by analyzing domain-specific glossaries.

2.3.2 Query Generation Methodology

- via a user study, for instance via questionnaires or focus group
- via a benchmarking analysis of existing sites
- heuristically based on the understanding of the domain developer a combination of the above

2.4 State of the Art

Assessing state of the art is one of the main important steps in the methodology. Here, we need to consult not only the existing resources related to the specific domain but also we need to scrutinize relevant standards, directives, guidelines, applications and ongoing projects. The main purpose of this step is to do the background study of the domain so that we can able to understand the current trends as well as current limitations which helps during modeling and implementation.

2.4.1 Standards

European Committee for Standardisation defined, a *standard* (French: norme, German: Norm) is a document that provides rules, guidelines or characteristics for activities or their results, for common and repeated use. Standards are created by bringing together all interested parties including manufacturers, users, consumers, and regulators of a particular material, product, process or service.

A *technical standard* is an established norm or requirement in regard to technical systems. It is usually a formal document that establishes uniform engineering or technical criteria, methods, processes and practices. In contrast, a custom, convention, company product, corporate standard,

etc. that becomes generally accepted and dominant is often called a *de facto standard*.

A *de facto standard* is a custom, convention, product, or system that has achieved a dominant position by public acceptance or market forces (such as early entrance to the market). Wikipedia define De facto is a Latin phrase that means in fact (literally by or from fact) in the sense of “in practice but not necessarily ordained by law” or “in practice or actuality, but not officially established”.

The *authority file* is a file or an authorized document maintained by librarians. This file is used to store and maintain the relevant information related to the books, such as author name, place of publication, etc. by using a distinct name value for each topic. The term authority file is widely used in the field of library and information science. The authority file is treated as an authorized document. These days base registries are controlled and maintained by European Public Administration. The European interoperability framework for European public services (EIF) maintains base registries, which provide authentic sources of basic information on items such as persons, companies, vehicles, licenses, locations, buildings and roads [Commission, 2010].

2.4.2 Resources

Resource is a source of help or information [Press, 2015]. In this step we need look for resources from which we can get terminology needed for representing concept of a given domain. Mostly, it is available from Language Resource. The term “Language Resource” refers to a set of speech or language data and descriptions in machine readable form, used for building, improving or evaluating natural language and speech algorithms or systems, or, as core resources for the software localization and language services industries, for language studies, electronic publishing, international

transactions, subject-area specialists and end users [Association, 2017]. For example, WordNet [Miller and Fellbaum, 1998] is a very famous general purpose language resource.

Other resources are rather different from language resources but they also treated as a source for terminology. They are subject specific control vocabulary, glossary, or local dictionary specifying meaning of the term use by local application or database.

2.4.3 Applications

Application is a program or piece of software designed to fulfil a particular purpose [Press, 2015]. In this step, we need to look for currently available applications which are brought use by organization or institute. This way we are able to analyze limitation and advantages for those existing application. The main objective should be to note down advantages from those applications and to adopt or adapt some functionality if possible. This way we are able to propose a modified version of the system architecture needed to be developed.

2.5 Informal Modeling Phase

As Subrahmanian et al. rightly stated in their paper [Subrahmanian et al., 1993] that formal equation are not enough if we are not consider informal model during the design process. Hales, S. [Sargent et al., 1992] find out that different modeler use different vocabularies to describe the same or very closely related set of things. Study showed that modeler typically spend at most 15% of their timing doing analytical task, the rest of their time being spent discussing various aspect of the modeling. This discussion most often taken the form of one-on-one meetings, discussion [Hales, 1987]. All of these informal aspects of modeling in the methodology need to be

considered otherwise being neglected.

2.5.1 Pilot Reference Datasets

This is also a vital step before start the modeling. In this phase, we need to verify local dataset of the given domain for which we want to create the model. It is to cross check, justify and scrutinize scope of the model with real datasets. As it is impossible to check all datasets of that domain, we need to select few datasets, referred to as pilot reference datasets from a different category to understand variety, similarity, and dissimilarity of those datasets.

2.5.2 ExER Model

A second part of the modeling was performed using the extension of the Entity-Relationship (ExER) model proposed by Peter Chen [Chen, 1976]. An Extended Entity-Relationship model is a conceptual (or semantic) data model, capable of describing the data requirements for a new information system in a direct and easy to understand graphical notation.

2.6 eTypes Model

In the past decade, ontologies have been used as core in most knowledge-based applications [Kharbat and El-Ghalayini, 2008]. In the literature, several definitions of ontology are available. Among them the probably most relevant definition of ontology was proposed by Guarino (1998): *a set of logical axioms designed to account for the intended meaning of a vocabulary* [Guarino, 1998]. In this definition, Guarino emphasized the role of logic as a way of representing an ontology. We believe that ontology has an important role to play in the general task of managing diverse information.

In particular, ontology can ensure coherent and correct conceptualization of the real-world entity providing the subject matter of the information to be handled. For example, road, highway, path, route are often used more or less interchangeably but they can have different intended meaning using ontology in the model help to minimize this confusion.

To model multivariate data, we choose an entity-centric approach to collocate all information in one place. We group real world entities as sets of *Entity types* (or in short *eTypes*). FGDC (Federal Geographic Data Committee) defines *eTypes* as “the definition and description of a set into which similar entity instances are classified (e.g. bridge)” [Committee, 2015]. An *eType* provides a schema and set of rules for the creation of a conceptual representation of a real world entity (e.g. a person, a building, an organization). We define an *Etype* as the quadruple,

$$eType = (ID, EC, NS, \{AD\}) \quad (2.1)$$

Where,

ID is a unique identifier;

EC is a concept denoting the class of the Etype;

NS is a name of the Etype;

AD is a non-empty set of Attribute Definitions.

AD determines the set of attributes that can be associated to instance of a certain *Etype*. An Attribute Definition is a tuple,

$$AD = (ID, AN, DT) \quad (2.2)$$

Where,

ID is a unique identifier (Here we are using ID corresponding to concept available in the vocabulary);

AN is the concept denoting the attribute name;

DT is a DataType.

In this context, we take the class of an *Etype* to be the most specific class which can be used to describe a specific instance of an *Etype*. An entity can only have one class. Thus for instance “pizzeria” and “restaurant” could be two classes for two entities of *Etype* building.

The full meta-model of the *eType* is illustrated in Figure 2.3. Notice that the *eType* name:NS (e.g. building), class:EC (e.g. restaurant, government building), attribute definition:AD (e.g. height, date of construction, roofing material) and qualitative attribute:QA are connected with the concept.

$$\text{Concept} \Rightarrow \{\text{EC}, \text{NS}, \text{AD}, \text{QA}\}$$

Notice also that a concept has a *semantic relation* (e.g. is-a, part-of, component-of) with its parent/child concept (e.g. building is-a structure). Concept also used for *synset*. *Synset* (i.e. sets of cognitive synonyms) contains terms (e.g. building, edifice) associated with the particular concept. A *Lexical relation* (e.g. synonym) show the relation between terms within a *synset*. *Semantic lexical relation* (e.g. hyponym, hypernym) is used to denote relation between synset. *Gloss* provides natural language description (e.g. building: “a structure that has a roof and walls and stands more or less permanently in one place”) of the concept. It helps to eliminate issues related with *heterogeneity in meaning*.

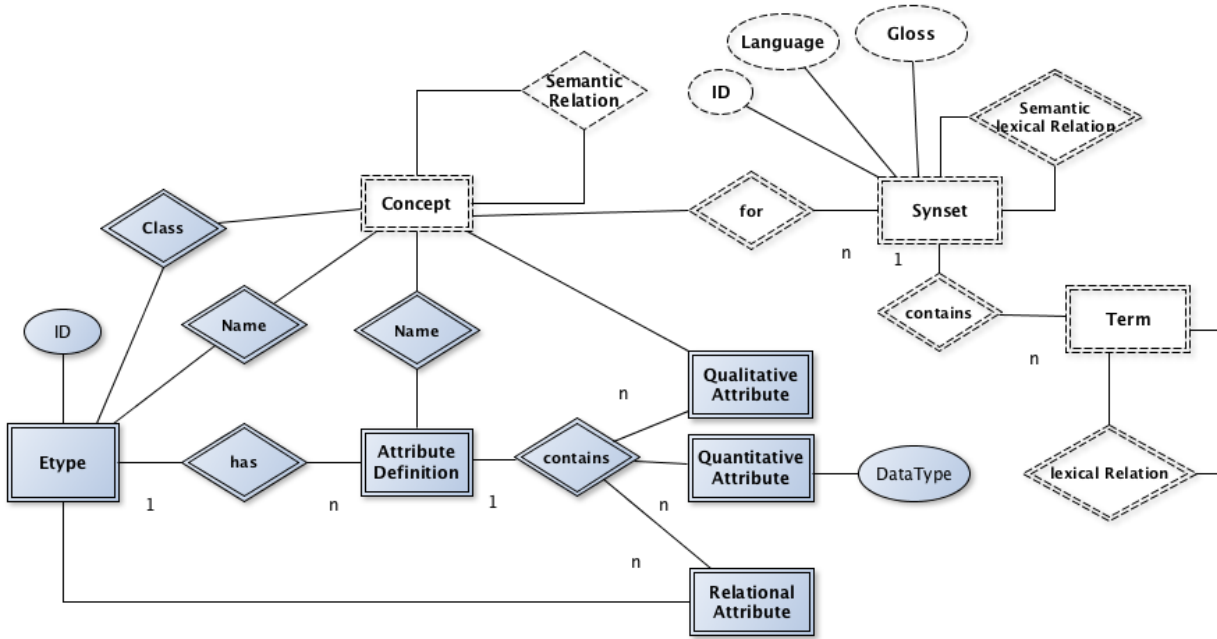


Figure 2.3: eType Metamodel

Notice that the meta-model clearly structures the *eTypes* model in two parts: (i) the one part centered on the *eType* which defines how to structure the schema. (ii) the second part centered around the concept which defines how to structure the vocabulary. Notice how the two parts are coupled via “concept”. In fact, notice that concept is used to capture and link to all the linguistics elements of the schema. This gives maximum flexibility in adaptivity and to all the terminology described. Hence, for instance, a word in one language schema can be represented in another language specified by vocabulary. In fact, to capture all simple terms used in the schema.

2.7 Schema Level

The schema level deals with the formal ontological description of the schema. It is the language independent part of the schema. We proposed the fol-

lowing steps to build the model.

1. **Defined Class:** In this step we need to organized all those concept which represent “A set or category of things having some property or attribute in common and differentiated from others by kind, type, or quality” [Press, 2015].
2. **Alignment with Upper Ontology:** Top/Upper-level ontologies describe very general concepts and provide general ” [Press, 2015] notions under which all root terms in existing ontology should be linked. For example, BFO [Arp and Smith, 2008], DOLCE [Guarino, 1998], UFO [Guizzardi et al., 2008], SUMO, CYC, YAMATO. For our work we uses DOLCE top-level [Guarino, 1998].

Top-level classes are:

Physical object: a concrete object that exists in space and time, especially one perceivable through the senses [Press, 2015]. For example, geographical location, building, person, artifact all falls under this category.

Artifact: man-made object such as Building, construction, car, bus, train are grouped under this top category. Building and construction belong to immovable objects, which has a permanent position on the Earths surface while car, bus and train belong to movable object.

Event: a thing that happens or takes place, especially one of importance [Press, 2015]. For example, any social gathering like festival or party, any kind of trip, even natural phenomenon such as avalanche, earthquake, or topical storm falls under this category.

Stative: expressing a state or condition rather than an activity or event, such as be or know, as opposed to run or grow [Press, 2015].

Any pathological state such as disease, trauma or natural phenomenon such as weather condition fall under this top category.

Mental object: the sum or range of what has been perceived, discovered, or learned. For example, book, prescription, note are group under this top category.

Person: a human being regarded as an individual [Press, 2015]. Person plays as an agent who usually builds connection among other entities. Person always participates in various events. One person can play different role in different contexts without changing his/her own identity.

Social object: objects that gain meaning through processes of reification (e.g. ritual). Studies of this phenomenon have its origins in classical cognitive sociology, the historical traditions of the sociology of knowledge and phenomenology ¹. For example, any social group formally created by law or informally created such as organization, company, team or government falls into this category.

3. Defined Attribute:

- Data type means a descriptor of a set of values that lack identity, in accordance with ISO 19103. Apart from standard data types, we also support others types such as:
 - NLString: Natural language string (NLString) allows the assignment of a String in a natural language.
 - SString: Semantic string (SString) allows the assignment of a semantic enabled value with semantics (possibly) computed from a string in a language.

¹https://en.wikipedia.org/wiki/Social_objects

- Concept: A special case of SString, where the value is exactly one Concept.
 - Entity: A special case of SString, where the value is exactly one Entity.
 - Complex type: A structure attribute. It is formed by nesting composite attributes and multi-valued attributes in arbitrary way ².
- Code list means an open enumeration that can be extended.
 - Enumeration means a data type whose instances form a fixed list of named literal values. Attributes of an enumerated type may only take values from this list.
 - External object identifier means a unique object identifier which is published by the responsible body, which may be used by external applications to reference the spatial object.
 - Identifier means a linguistically independent sequence of characters capable of uniquely and permanently identifying that with which it is associated, in accordance with EN ISO 19135.

4. **Defined Relation:** Relations are used in order to connect entities among each other. For example, *part-of* relation is used between City and Country or between Suburb and City. This is used to say, for example, that Povo is part of Trento and Trento is part of Italy. Similarly the relations *addressCountry*, *addressCity* and are used to define the geographical position of a building.

²http://databasemanagement.wikia.com/wiki/Category:Complex_attribute

2.8 Language Level

In this part, the lexicalization in multiple languages is provided for all the concepts (denoting entity class, relation, and attribute). Within each language, terms with the same meaning are grouped together into sets of synonyms, called synset (e.g. building, edifice). A synset is defined as,

$$\text{Synset} = (\text{ID}, \{\text{ST}\}, \text{R}, \text{G})$$

Where,

ID is a unique identifier;

ST is a set of synonymous terms;

R is a relation among synsets;

G is a glossary (i.e. natural language description)

This level can be instantiated in multiple languages (e.g., in English and Italian). For example, Disease is defined as an impairment of health or a condition of abnormal functioning in English or *Malattia* is defined as *un indebolimento della salute o un funzionamento non normale* in Italian are the same concept in different language. This type of separation ensures high flexibility in the way data can be described and communicated. This level clearly addresses *diversity in language*. To develop the linguistic part we applied the library science principles [Ranganathan, 1967] in different steps of our process.

Mapping between Domain-specific Language and concept with in the our proposed framework represented in Figure 2.4. Here, the English word breast cancer has two meanings, as general meaning and as domain-specific meaning, which are represented by two single word synsets. General meaning i.e. *breast cancer* (in English) and *cancro al seno* (in Italian) are understand by all common people. Whereas domain-specific language i.e ICD10:C50³ also referring the same concept of *breast cancer*.

³International Classification of Diseases (ICD) version 10 code for Breast cancer

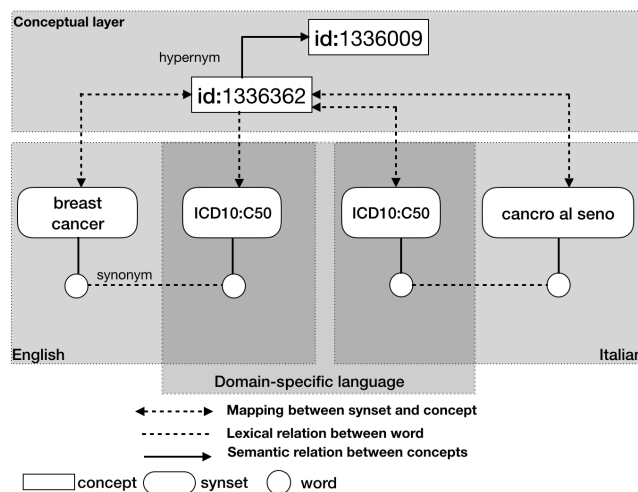


Figure 2.4: Relation between schema level and language level

2.9 Evaluation

Evaluation is a continuous process. For some part of the methodology we need to perform evaluation through out the process, as they are rather implicit and need to be checked continuously. For example, ontological commitment and quality of the model. Others evaluation can be verified only after completion of the model. In our methodology we followed three criteria describe below.

2.9.1 Model Evaluation

For evaluation of the model, we followed the guideline as proposed by Gmez-Prez [Gómez-Pérez, 2001]. According to Gmez-Prez [Gómez-Pérez, 2001], the goal of the evaluation process is to check what the developed ontology defines correctly, does not define, or even defines incorrectly. Two steps needed to be followed and they are: verification and validation. The purpose of verification is to check the syntactic correctness. The purpose of validation is to check its consistency, completeness and conciseness. Ontology editors, such as Protégé, typically provide facilities to check syntactic

correctness and consistency can be checked by the reasoner such as Hermit⁴, Fact++⁵ which are available as a Protégé plugins. The model is complete if it fully captures what it is purported to represent of the real world. The model is concise if it does not accommodate redundancies. We ensure that the developed model is complete and concise by inducing the necessary entity types and properties from the competency questions.

Quality of the model from ontological commitment point of view has been checked according to *OntoClean* methodology suggested by Guarino and Welty [Guarino and Welty, 2002]. The OntoClean methodology is a domain independent, based on formal notions, which are general enough to be used in any ontology effort. They are:

Essence and Rigidity: The first formal criteria needed to be check is essence. A property of an entity is essential to that entity if it must hold throughout the time. For example, consider the property pathological condition. We may say that it is an essential property of patient, but not of person. A special form of essence is rigidity. A property is called rigid if it is essential to all its individuals. For example, being a person is usually conceptualized as rigid, while we have seen that being patient is not.

Identity and Unity: Identity criteria is a notion to check whether two entities are the same entity or not. This is a difficult task even for an experienced conceptual modelers to appreciate because they are typically not part of the implemented system and are overlooked. For example, whether bus stop X and bus station X is a same entity or different. This is one of the common confusions of natural language when used in describing the world. When we say “all bus station are bus stop we really mean “all bus station have a bus stop. Other notion is Unity. Unity helps to find the

⁴<http://www.hermit-reasoner.com>

⁵<http://owl.man.ac.uk/factplusplus/>

intended meaning of properties or classes based on whether their instances are wholes. For some classes, all their instances are wholes, for others none of their instances are wholes. For example, the class “step, found in some commonsense ontologies, does not represent whole objects. An instance of this class is step of a building, but it is not a whole, since it is not recognizable as an independent entity. An other example, could be “ward” of a hospital which should not be treated a whole.

Subsumption Misused: the most unclear matter is to distinction between the two relations subsumption and instantiation. There are many example of subsumption relationship misused when instantiation was actually intended. For example, the difference between building and location. Building has location but it does not means that building is a subclass of location, like we see in many geographical information system (GIS) based applications. So in case of instantiation a building does not inherit all properties from those of location like elevation, bio-geographical characteristic or weather condition.

Part/Whole: Sometimes, it is difficult distinguish between the part-of and the sub-class as subclass is analogous to subset, and a subset of a set is a part of it [Guarino and Welty, 2002]. However, it is often notice misuse subsumption to represent part-of during implementation of the model. This needed to be check before implementation. For example, department of a university or ward of a hospital are related with part-of relation and not subsumption.

Polysemy: Problem with natural language is that it always has multiple meaning and human can able to interprets different meaning depending on the situation or context. For example, the term cold can be referred to the common cold or low or relatively low temperature. During modeling we need to aware of this kind of situation whether use a term as a class or attribute value.

Checking the ontological commitment according to *OntoClean* is an implicit process. We ensure these above mentioned criteria by taking part of the class lecture from Guarino and forward the same during the KDI⁶ course. In addition, we verify its empirical adequacy by checking that information extracted from popular datasets or websites in the domain can be comfortably represented by the developed model.

2.9.2 Terminology Evaluation

Terminology evaluation is the task to check whether all terms required to capture knowledge for the model including class names, attribute names, and also attribute values are present in the knowledge base. Best practical suggestion would be to include the language resource which has a large control vocabulary. For example, in case of Healthcare domain we imported the entire SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms) [International, 2017] vocabulary which is the largest standard vocabulary for medical and healthcare. Other process could be import many resources monotonously and then do the statistical analysis to check the coverage.

2.10 Case Studies

We successfully implemented our methodology in three application domains. Detailed discussion can be found in latter chapter of the thesis.

2.11 Related Work

We categorized related work into two types. In the first part, we gave a brief overview all relevant ontology development methodology. Most

⁶<http://disi.unitn.it/ldkr/ldkr2016/lessons.html>

of them are actually been developed based on the experience on a single domain. So it is necessary to consider them as part of domain development methodology. In the second part of the related work, we described work on ontology evaluation techniques.

2.11.1 Ontology Development Methodology

The literature is overloaded with studies in which several researchers have proposed various ontology development methodologies. Here we mentioned few of them from which we got inspiration to develop our proposed methodology and understand the gaps/flaws in them.

IDEF5 Capture Method: IDEF5 project provides a set of guidelines to various knowledge engineers, analysts and researchers to develop and manage ontologies efficiently [Benjamin et al., 1994].

Uschold And King's Methodology: This methodology has been proposed based on the experience of developing the Enterprise Ontology. This is the first methodology of this kind which provides guidelines for developing ontologies, they are: 1) Identify the purpose and intended uses for the proposed ontology. 2) Second phase is building the ontology itself, which again divided into three sub steps, they are: i) identification of the key concepts and relationship in the given domain. Identification of terminology for the concepts and relationships. ii) In coding phase deals with the explicit representation of the knowledge acquired in the previous steps. iii) In integrating process, there is the question of how to use existing ontologies. 3) Third phase is evaluation to make a technical judgement of the ontologies. It can be check against requirements specification, competency questions, or the real world [Gómez-Pérez et al., 1995]. 4) Finally, documentation for the whole ontology process and possibly writing down all guidelines followed and justification in case of differences.

Gruninger And Fox's Methodology: Gruninger and Fox proposed

a methodology using first order logic which is inspired on the development of knowledge-based systems. This methodology has been suggested as TOVE (TOronto Virtual Enterprise) project ontology within the domain of business processes and activities modeling. This represents logical model of knowledge. The steps follows in the methodology are: 1) Describing the motivating scenarios. 2) Formulation of informal competency questions, to set the scope of the planed ontology. 3) Formulation of formal competency questions, which specify the terminology with definition and constraints. 4) Specification of axioms and definition within the formal language. 5) Finally, specify the conditions under which the solutions to the questions are complete. In this methodology, the ontology can be create by using questions and answers for predefined motivating scenarios, which represents main concepts, properties, relations and axioms on the ontology [Gruninger, 1995, Fernández-López, 1999] . The methodology is very well structure and can be extend the scope.

METHONNTOLOGY Methodology: This methodology proposed a structured method to build ontologies from scratch [Fernández-López et al., 1997]. Seven steps are followed in the methodology they are 1) Specification phase, 2) knowledge acquisition, 3) Conceptualization 4) Integration, 5) Implementation, 6) Evaluation and 7) Lastly, documentation.

SENSUS Methodology: SENSUS-based methodology was proposed while developing the SENSUS ontology at the ISI (Information Science Institute) natural language group to provide a broad conceptual structure for developing automated machine translator [Knight et al., 1995]. Five steps are taken to build a particular domain ontology, they are: 1) First, collect a series of terms are taken as *seed*. 2) These seed terms are then linked manually to SENSUS. 3) All collected concepts in the path from the seed to the root of SENSUS need to include. 4) Finally, to check those nodes which have a large number of path and add the entire sub-tree

if needed. This is also a manual process, since it seems to require good understanding of the domain to make the right decision [Swartout et al., 1996].

WordNet Methodology: WordNet is a lexical database for the English language. It groups English words into sets of synonyms called synsets, provides short description of all terms, and captures the various semantic relations between these synonym sets [Miller and Fellbaum, 1998]. The purpose of WordNet is twofold: to produce a combination of dictionary and thesaurus that is more usable, and to support automatic words sense disambiguation and natural language processing task. The hypernym/hyponym relationships among the noun synsets can be interpreted as specialization relations between conceptual categories. In other words, WordNet can be interpreted and used as a lexical ontology in the computer science. However, such ontology should be corrected before being used since it contains hundreds of basic semantic inconsistencies such as (i) the existence of common specializations for exclusive categories and (ii) redundancies in the specialization hierarchy. Furthermore, transforming WordNet into a lexical ontology usable for knowledge representation should normally also involve (i) distinguishing the specialization relations into *subtypeOf* and *instanceOf* relations, and (ii) associating intuitive unique identifiers to each category.

Ontology Development 101: This methodology proposed by the Stanford Center for Biomedical Informatics Research (BMIR) to develop ontology using Protégé tool [Noy et al., 2001]. The wine and food example were used in the methodology guide, which is loosely based on an example knowledge base presented in the paper on a description-logics approach by Brachman et al. [Brachman et al., 1991]. However, methodology also clarify that ontology development is different from designing classes and relations in object-oriented programming. This guide tried to provide de-

veloper an initial road-map that would help a new ontology designer to develop ontologies. Finally, authors concludes that there is no single correct ontology-design methodology and this methodology should be use as a references for domain ontology development.

Integrated Methodology: Integrated ontology development methodology [Chaware and Rao, 2010] is quite similar with the Gruninger and Fox methodology [Gruninger, 1995] and divided into four steps: 1) Motivating user scenarios or keyword. 2) Formulation of informal/formal questions and answer module. 3) extracting of terms and constraints module. 4) Finally, build ontology based on top-down approach. This methodology is validated only on shopping mall scenario.

Yet Another Methodology: YAMO is a Yet Another Methodology for large-scale ontology development. The methodology is motivated by facet analysis and an analytico-synthetic classification approach. The approach ensures the quality of the system precisely inter of its flexibility, hospitable, extensible, dense and complete. YAMO consists of two-way approaches: top-down and bottom-up. YAMO food as an example domain and use that domain to defined their ontology. A user interviewed has been conducted with a group of people to gather a practical overview, which provided more insight into the theoretical understanding of the domain [Dutta et al., 2015].

Limitations of Existing Ontology Development Methodologies: Existing methodologies do have some limitations. It is briefly describe in the work of Chaware and Rao [Chaware and Rao, 2010]. The limitation are:

- Some of the methodologies are very formal and fit for small-scale applications or contexts.
- Some methodologies like Methontology, is better structure and de-

tailed where as some steps can be either adopt or adapt depending on context [Öhgren and Sandkuhl, 2005].

- Full integration of existing ontologies may be complicated due to change in structure or sequence.
- For each and every scenario it is most likely not possible to decide the competency questions, which will represents the rules and constraints of terms used in model.
- No clear guideline on how to use existing standards.
- There is a lack of clarification on how to use and extract the terms. No details exist on how to deals with synonymous words and how to handle codification system used in some specific domain application.
- Existing methodologies are never tested or evaluated from the user experience (UX) point of view. That means that it has never be measured from user perspective i.e., whether a methodology is easy to understand, practical training is required or not etc.

2.11.2 Ontology Evaluation Technique

The evaluation of ontologies is an ongoing research field. At present, there is quite a few number of techniques available. Here, We enlisted evaluation criteria defined by five important papers from literature. These quality criteria need to be followed as prerequisite, goals to guide the creation and evaluation of the ontology. Asunción Gómez-Pérez lists the following criteria [Gómez-Pérez, 2001]:

- **Consistency:** Capturing both the logical consistency (i.e. no contradictions can be inferred) and the consistency between the formal and the informal descriptions (i.e. the comments and the formal descriptions match)

- **Completeness:** All the knowledge that is expected to be in the ontology is either explicitly stated or can be inferred from the ontology.
- **Conciseness:** Whether the ontology is free of any unnecessary, useless, or redundant axioms.
- **Expandability:** refers to the required effort to add new definitions without altering the already stated semantics.
- **Sensitiveness:** It is related to how small changes in an axiom alter the semantics of the ontology.

Thomas Gruber defines the following criteria [Gruber, 1995]:

- **Clarity:** An ontology should effectively communicate the intended meaning of defined terms. Definitions should be objective. When a definition can be stated in logical axioms, it should be. Where possible, a definition is preferred over a description. All entities should be documented with natural language.
- **Coherence:** Inferred statements should be correct. At the least, the defining axioms should be logically consistent. Also, the natural language documentation should be coherent with the formal statements.
- **Extendibility:** An ontology should offer a conceptual foundation for a range of anticipated tasks, and the representation should be crafted so that one can extend and specialize the ontology monotonically. New terms can be introduced without the need to revise existing axioms.
- **Minimal encoding bias:** An encoding bias results when representation choices are made purely for the convenience of notation or implementation. Encoding bias should be minimized, because knowledge-sharing agents may be implemented with different libraries and representation styles.

- **Minimal ontological commitment:** The ontology should specify the weakest theory (i.e. allowing the most models) and defining only those terms that are essential to the communication of knowledge consistent with that theory.

Chapter 3

Geospatial Domain

“You can design and create, and build the most wonderful place in the world. But it takes people to make the dream a reality.” — *Walt Disney*

3.1 Domain Description

Without maps we would not even know where we are. Geography matters in our day-to-day life. It affects the university we went to, the mother tongue we speak, who are our neighbor countries, our daily movement, etc. In a single sentence, geography impacts almost everything. We mostly use some type of geographic data starting from the old day paper maps to maps in a cell phone. We use maps to find where we are going. Data with a geographical component (i.e. longitude, latitude or weather) connected to some place on the earth called as a spatial data [Nation, 2013]. We use spatial data to find hotel, population, and country to visit. Maps are everywhere we look. They help us to learn our World and with the help of spatial data, we can assess the global issues such as climate change, risk from flooding, examine the spatial distribution of nation populations and

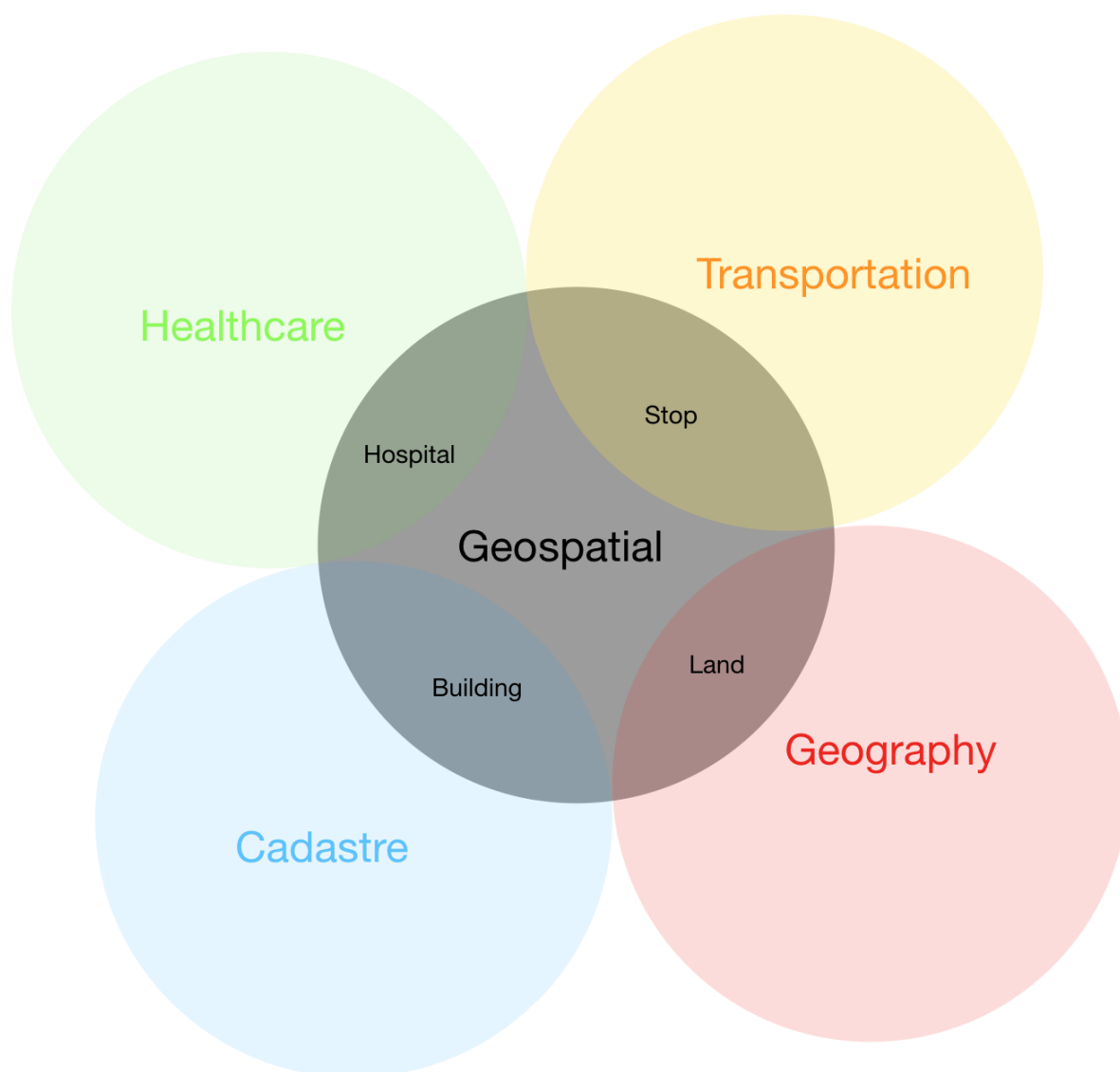


Figure 3.1: Geospatial Domain

natural resources or best place to business, how much green space in the city, so on and so forth. In our domain when we use spatial data then we means geospatial data (data refers to places on the earths surface at human scale (unit of measures for length, size and volume etc.). Using spatial data we can model real world systems and their interactions. Using spatial analysis, we can model and visualize our World. The use of geospatial information is rapidly increasing. There is a growing recognition by both government and the private sector that understanding of location and place is a key element of fruitful decision making [Nation, 2013]. The terms location and place in geography are used to identify a point or an area on the Earth’s surface or elsewhere. The term location generally implies a higher degree of certainty than place, which often indicates an entity with an ambiguous boundary, relying more on human/social attributes of place identity and sense of place than on geometry according to Wikipedia. Oxford English dictionary define Location as *A particular place or position* and gazetteer usually contain information about location instance. In existing geospatial resources [Giunchiglia et al., 2010] and even in INSPIRE directive [Directive, 2007] has a grey area where they do not specify what are the entity should be consider as geospatial entity. Figure 3.1 illustrated how geospatial domain intersects with other domains using the set diagram. In our document we explicitly mention this thing. In the Word-Net feature of earth distribute under five main domain namely i. Geology ii. Meteorology iii. Oceanography iv. Paleontology and v. Geography [Bentivogli et al., 2004]. We are considering location, which belongs to geography field.

3.1.1 Basic Definition

Here we listed few definition of the terms used in this chapter. The usefulness of this definition is to clarify the meaning of the term explicitly for

avoiding (or limiting) further confusion.

Space: Ranganathan, [Ranganathan, 1967] and Giunchiglia et. al. [Giunchiglia et al., 2010] define space as *We consider Space in accordance with what people commonly understand by this term, which includes the surface of the earth, the space inside it and the space outside it.*

Location: Location has different meaning in different context. Word-Net¹ definition is very generic in nature. It defines location as *a point or extent in space*. Wikipedia defined Location as *the terms location and place in geography are used to identify a point or an area on the Earth's surface or elsewhere. The term location generally implies a higher degree of certainty than place, which often indicates an entity with an ambiguous boundary, relying more on human/social attributes of place identity and sense of place than on geometry*; Location is *a particular place or position* as per oxford English dictionary. Merriam-Webster dictionary defined location as *a place or position*.

Spatial object: *an abstract representation of a real-world phenomenon related to a specific location or geographical area* [Directive, 2007] this term synonymous with the ISO 19100 term *geographical feature*.

Spatial entity: *an entity that is not inherently a location, but one which is identified as participating in a spatial relation is tagged as a spatial entity* [Pustejovsky et al., 2011].

Spatial things: *anything with spatial extent, i.e. size, shape, or position. E.g. people, places, bowling balls, as well as abstract area like cubes* [Brickley, 2004]. spatial things are that subset of real-world phenomena which related to a location [Directive, 2007].

Geo-referencing: When we describe any phenomena or object in respect with the geography (i.e. Earths surface) then that process is called Geo-referencing.

¹<https://wordnet.princeton.edu>

Geo-coding is the process of finding a geographic location (i.e. geo-coordinate) based on other information, like a street address, postal code etc.

Reverse geo-coding is the process of back coding of a point location (e.g. longitude, latitude) to a human readable address or place name like a street address, city.

Geospatial entity: The first question is what is meant by a Geospatial entity. The answer is, there is no universally accepted definition of a geospatial entity. It means different things to different people [Nation, 2013]. The conceptualization of geospatial entity, therefore, varies from country to country, researcher to researcher, depending on the context and application, scenario and purpose of the application, potential users for whom the application is proposed. A geospatial entity would have a different connotation in our application domain than, say, Geographical Information System (GIS) [Baglioni et al., 2011]. Even in GIS, there is no one-way of defining a geospatial entity. For our application we define geospatial entity as *a physical object (here physical object means a tangible and visible physical entity), which has an existence in our planet Earth (aka World) and occupies certain geometry area which represent as point, line, and polygon, and maintains their identity through time.*

In the real world representing all geospatial entities is not same. Some geospatial entities have well-define boundaries (a building, a road) and some have fuzzy boundaries that partially depends on human cognition (a city, a mountain, a bay etc.) [Smith and Varzi, 2000]. Entities do not have any spatial (i.e. Geographical) and temporal attributes are outside the scope of our work. We excluded concept related to space research such as outer space, interstellar space; metaphoric concept related to space for example Hell on earth, bilocation; economic concept related to space such as banana republic, tax haven. Movable object also excluded from our

scope of work.

3.1.2 Purpose and objectives

The fact itself that we base the work on INSPIRE [Directive, 2007] will guarantee the relevance and usefulness of the model. We expect our entity-centric model to be of higher quality than state of the art solutions such as YAGO3 [Mahdisoltani et al., 2014]. We will compare these resources or approaches in terms of quality of metadata and terminology. We will validate the work using standard ontology evaluation techniques. We expected this to enable interoperability in the domain.

3.1.3 Problems

Existing geospatial resources mainly have three problems. (a) Lack of terminological support: without proper meaning of the terms. It is impossible for the common people to understand relation between concepts. For example, same term might have different meanings. Sometimes meaning of the term court can be *A body of people presided over by a judge, judges, or magistrate, and acting as a tribunal in civil and criminal cases* [Press, 2015] or *A quadrangular area, either open or covered, marked out for ball games such as tennis or squash* [Press, 2015]. The definition of proper standard vocabularies that describe Geo-spatial information is one of the major challenges in geospatial data management as pointed out by the GeoKnow3 project [Le Grange et al., 2014]. So it is necessary to provide meaning of the terms not only for human consumption as well as machine consumption [Baader, 2003]. (b) Lack of temporal information: entities come into existence at a particular time point and disappear after a particular time point [Wang et al., 2010]. For example, ancient sites are revealed and concealed. Capital of a country may not be always permanent. Buildings are

constructed and destructed. (c) Lack of attributes: for addressing domain specific scenario we need more attributes, which are not available in a single dataset. There is also problem of data quality in the existing resources. Problems (a) and (b) are applicable for all geo-spatial entities in general. Problem (c) is applicable only for specific scenarios mentioned above.

Concerning the mentioned problem manually built geospatial resources like GeoWordNet [Giunchiglia et al., 2010] provides terminological support but does not provide any temporal information of the geospatial entities. It is not built in any domain specific applications. GeoNames and OpenStreetMaps (OSM) are famous resources for the Geo-spatial application but they do not take into account the problem (b). GeoNames provides a description (e.g. gloss) for most of its categories (e.g. features class). OSM provides tags for categorizing the entities and specify their intended use in their Wiki pages⁴ [Codescu et al., 2011]. This pertains problem (a). On the other hand, auto generated Geo-spatial ontology, such as T-YAGO [Wang et al., 2010] and YAGO2 [Hoffart et al., 2013] which considered both spatial and temporal aspects but they have several limitations for instance: syntactic matches, match one word with another word which have the same syntax like *Times_New_Roman* with *New_Roman_Times* assuming same entity whereas first one is typo introduce by the Times newspaper and second one is name of a music album released in 2004. Quality of data provides by YAGO2 also questionable.

3.2 Reference Scenarios

Geospatial domain is applicable in very vast number of applications. They are:

- Tourism (Cultural heritage tourism, Mountain tourism, Agri-tourism)
- Journey planner

- Traffic management
- Waste management
- Utility management (aka government service or public service)
- Postal collection or delivery management
- Urban planning (aka city planning)
- Real estate management (aka land administration)
- Forest management (aka conservation management)
- Emergency management (aka risk/disaster management)
 - Flood management [After emergency]
 - Early warning system [Before emergency]
 - Hazardous material management [e.g. Nuclear power plant]

3.3 Generalized Queries

Give me X geospatial entity with time constraint t , attribute constraint Y and location constraint Z

According to the UNWTO [Nation, 2016], Europe was the most popular visited continent in the world in 2013. The cultural heritage of Europe, the beautiful landscapes and the quality of its tourist establishment is the main reasons why tourists choose to take their vacations in Europe. Main factors of selecting accommodation by the tourist are location, purpose of the uses (e.g. business or pleasure), current needs (e.g. honeymooners, families with children or religious convention), price [Gagnon and De Souto, 2016].

Question generally ask in the contexts of tourism will fit into the generic query schema provided above. Example of few frequently ask queries are:
Give me

- Q1 all the World heritage sites (X) which got enlisted in 2014 (t) from country (Z)

- Q2 all the five star (Y) hotels (X) located in city (Z)
- Q3 all the hotels (X) price (Y) within \$100 per night in city (Z)
- Q4 all the hotels (X) that has parking facility (Y1) and swimming pool (Y2) in city (Z)
- Q5 all the rented accommodation (X) price (Y) within \$200 near the costal area (Z)
- Q6 all the hotels (X) that have more than 30 bedrooms (Y1) and room rate within \$150 (Y2) located in city (Z)
- Q7 all the hotels (X) that have more than 100 rooms (Y) located in city (Z)
- Q8 all the hotels (X) that have saunas (Y) located in city (Z)
- Q9 all the economy (Y) hotel (X) near the airport (Z)
- Q10 all the resort (X) that have spas (Y) located near spring (Z)
- Q11 all the budget (Y) lodge (X) that located in or near national park (Z)
- Q12 all the agritourism (X) that have educational activities (Y) for children in Town (Z) Q13 all railway stations (X), which started in early 20th century (t) in city (Z)
- Q14 all the museums (X) that is open now (t) located in city (Z)
- Q15 all the theatres (X) that is open now (t) located in town (Z)
- Q16 the best quality (Y) bathing water lake (X) in at most two hours by road (Y)
- Q17 the best city (X) in terms of air quality (Y) in continent (Z)
- Q18 the ski resort (X) located in Mountain (Z) of the country (Z)
- Q19 the charging station (X) that is open now (t) located in city (Z)

3.3.1 Identify the Domain Concepts

Some useful definition of geospatial Concepts are:

Space region: Geographical region separated by administrative purpose (e.g. Country, Region) or classify based on the earths vegetation pattern (Alpine region, savanna region, tundra region) and region filled with water (e.g. sea region) group under space region.

Feature: The real world object which can have a location treated as "feature" (or geographical feature) in all GIS based system [Perry and Herring, 2012].

Here we describe all the INSPIRE [Directive, 2007] themes which are related to the geospatial domain.

Administrative unit: *Units of administration, dividing areas where Member States have and/or exercise jurisdictional rights, for local, regional and national governance, separated by administrative boundaries* [Directive, 2007]. In the other word Geospatial region divided for administrative purpose. Administrative unit usually administered by only one administrative authority. For example, USA (country) administered by Government of the USA.

Cadastral parcel: *the cadastral parcel should be, as much as possible, single area of Earth surface (land and/or water) under homogenous real property right and unique ownership, where real property right and ownership are defined by national laws* [Directive, 2007]. Only register part of the Earths surface such as urban cadastral consider by the land registry. Mandatory elements of the cadastral parcel are, i) Geometry, ii) Unique ID, iii) cadastral reference (e.g. *agenzia del Territorio (Italy)*), iv) label of the parcels (available in printed maps e.g. label of the land parcel of POVO 1 and 2 building is 5 as per here maps used in all windows phone). Italian cadastral databases have two parts. Land cadastre and Building cadastre, which contain 83.5 million of rural parcels and about 60 million of real estate urban parcels respectively [Maggio, 2012].

Hydrography: *Hydrographic elements, including marine areas and all other water bodies and items related to them, including river basins and sub-basins* [Directive, 2007]. It has many users and uses.

Transportation network: Wikipedia defined “A transport network, or transportation network is a realization of a spatial network, describing a structure which permits either vehicular movement or flow of some commodity. Examples are network of roads and streets, railways, pipes, aqueducts, and power lines”. Every component of the transportation network is not belongs to location. Few elements are artifact such as Bus, Train and Taxi etc (aka non geo-locating object [Pustejovsky et al., 2011]. and some have been considered as a location. For avoiding any ambiguity, in our work we consider the transportation area limited to the location from where we can avail transportation service. It consists of two elements, namely transportation point (bus stop) and transportation link (such as railway lines, road etc.)

- Transportation point: bus stop, buoy, beacon etc.
- Transportation link: railway line, road.
- Transportation object : bus, train, cable car, ship etc.
- Transportation node: bus station, railway station, airport

Protected site: *Area designated or managed within a framework of international, Community and Member States’ legislation to achieve specific conservation objectives*[Directive, 2007]. International Union for Conservation of Nature (IUCN) ² is the main international body, which maintains a central database of world-protected site. For example, Yellowstone national park, USA; Parco Nazionale Dolomiti Bellunesi, Italy.

²<http://www.iucn.org>

Bio-geographical region: *Bio-geographical region describe areas of relatively homogeneous ecological conditions with common characteristics [Directive, 2007]. Based on vegetation cover European continent divided into nine category such as Alpine region, Arctic region, Mediterranean region.*

Habitats and Biotopes: *Geographical areas characteristics by specific ecological conditions, processes, and (life support) functions that physically support the organisms that live there. Including terrestrial, fresh water and marine areas distinguished by geographical, abiotic and biotic features, where entirely natural or semi-natural [Directive, 2007]. For example, woodland, nesting place etc.*

Natural Risk Zone: *Vulnerable areas characterized according to natural hazards (all atmospheric, hydrologic, volcanic and wildfire phenomena that, because of their location, severity, and frequency, have the potential to seriously affect society) e.g. floods, landslides and subsidence, avalanches, forest fires, earthquakes, volcanic eruptions [Directive, 2007]. Natural risk zone are zones where natural hazard areas are coincident with highly populated area and/or areas of particular environmental, cultural, or economic value. For example, Lombardy, Valsugana area consider as an avalanches zone in Italy.*

Building: *A building is a covered facility, usable for the protection of humans, animals, things or the production of economic goods. A building refers to any structure permanently constructed or erected on its site. Information on location of buildings may be supplied as points or with the actual basic form of the building. Usually buildings are part of cadastre. On the local level buildings are available within the large scale cadastral maps or cadastral data sets and are geometrically represented as surfaces [Directive, 2007]. Building can be used for different purpose such as, dwelling (e.g. house), commercial (e.g. office), agricultural (e.g. farm building) uses.*

Facility: Facility mainly provides services to the people (e.g. Railway station) and some facility help to get particular services (e.g. Weather monitoring station) all are grouped under the facility. *A Facility represent something designed, built, installed to serve a specific function, comprehending the complete equipment or apparatus for a particular process or operation. A facility groups together one or more installations that are operated on the same site by the same natural or legal person and, where present, the land, buildings, and equipment used in carrying on an industrial, business, or other undertaking or service* [Directive, 2007]. We consider only installing facility which has a permanent address on the surface of the earth for its identification. Some facilities are depending on building such as Hotel, Restaurant, which cannot run its service without building.

- **Production and industrial facility:** *This theme comprises features related to production and industry, as well as entities related to describing summary information about the activities taking place in Production and Industrial Facilities, and the main environmental issues related to them (pollution prevention, waste management, risk* [Directive, 2007]. For example, factory, industrial plant, production site etc.
- **Agricultural and Aquaculture Facility:** *Farming equipment and production facilities (including irrigation systems, greenhouses and stables* [Directive, 2007]. For example, pumping station, farm area, agricultural land, irrigation ditch, farm building etc.
- **Environmental monitoring facility:** *Location and operation of environmental monitoring facilities includes observation and measurement of emissions, of the state of environmental media and of other ecosystem parameters (biodiversity, ecological conditions of vegetation, etc.) by or on behalf of public authorities* [Directive, 2007]. For exam-

ple, the European environment agency (EEA) ³ installed monitoring station in different part of the Europe for monitoring weather, air-quality and emission.

- **Utility and Governmental Service:** *Utility and Government services includes utility facilities such as sewage, waste management, energy supply and water supply, administrative and social governmental services such as public administrations, civil protection sites, schools and hospitals* [Directive, 2007]. This theme provides basic information (e.g. the location, basic technical characteristics or involved parties) on a wide range of administrative and social services of public interest such as water supply, sewerage system, energy supply, communication network, maintenance of supply and vulnerability.

Some polyonymous concepts, we encounter during modeling are:

- **Condominium** “An administrative area established independently to any national administrative division of territory and administered by two or more countries” [Directive, 2007].
- **Condominium** “Housing consisting of a complex of dwelling units (as an apartment house) in which each unit is individually owned” [Miller and Fellbaum, 1998].

3.3.2 Query Collection

In principle, there are various ways to collect queries. For example, this can be done by analyzing the millions of user queries stored in the query logs of existing search portals or from interviewing the peoples. For the relevant queries, we consulted Travel career development handbook [Gagnon

³<http://www.eea.europa.eu>

and De Souto, 2016] which provided an in-depth discussion of all tourism aspect.

3.4 State of the Art

3.4.1 Standards

INSPIRE: INSPIRE is the European standard on Location[Directive, 2007]. The Directive aims to create a European Union spatial data infrastructure for the purposes of EU environmental policies and policies or activities which may have an impact on the environment. The Directive addresses 34 spatial data themes needed for environmental applications.

GeoSPARQL: GeoSPARQL is an emerging standard within the Open Geospatial Consortium (OGC)[Kolas and Batle, 2012]. Its intent is to provide a standard way to express and query spatial elements in RDF, so that users can exchange data easily, and triple store implementors can have a standard format for indexing. The purpose of this document is to provide an easy introduction to GeoSPARQL for Semantic Web users. This document assumes a working understanding of RDF and RDFS, and the Turtle serialization of RDF. This document is not intended to be exhaustive, but more topics and deeper explanations may be added over time.

GTFS: The General Transit Feed Specification (GTFS), also known as GTFS static or static transit to differentiate it from the GTFS realtime extension, defines a common format for public transportation schedules and associated geographic information[Google, 2006]. It is use as a *De facto standard* on Transportation. GTFS “feeds” let public transit agencies publish their transit data and developers write applications that consume that data in an interoperable way.

3.4.2 Resources

We address the state of the art from the point of view of systems such as, Geographical information system (GIS), Historical GIS (HGIS) and from resources such as, GeoNames [Wick and Vatan, 2012], GeoWordnet [Giunchiglia et al., 2010] and others. GIS community is a pioneer in Geo-spatial domain and continues their research on how to address the tasks in the context of a digital environment, combining models for better analysis of Geo-spatial information and visualization. GIS evolves from numeric cartography integrating remote sensing and digital images, typically skipping any conceptual design and modeling phase [Baglioni et al., 2011]. They are technologically very rich in terms of spatial analysis but do not provide temporal information. However, many HGIS provides temporal information and application-ready solution to visually represent the meaning of Geo-spatial information but did not provide any terminological support for the terms used in their system. We compared existing HGIS on the basis of support queries or not, time, the schema (ontological model) and terminology (vocabularies describe the meaning of the term) and we found that nobody provides the terminological support and also do not have any schema according to best of our knowledge.

GeoNames [Wick and Vatan, 2012] is a rich source of information about Geo-spatial entities and contains over 10 million geographical names and over 9 million unique features. It contains geographical data such as place names in multiple languages, latitude, longitude, altitude and population collected from different data sources including Wikipedia. It has total 645 categories (e.g. features class) classified under 9 top categories. It has total 645 feature classes and gloss has been provided for most of its features code, but there is no explicit meaning of the feature code. Sometime two different codes have the same meaning, for instance (AIRP) airport

(a place where aircraft regularly land and take off, with runways, navigational aids, and major facilities for the commercial handling of passengers and cargo) and (AIRF) airfield (a place on land where aircraft land and take off; no facilities provided for the commercial handling of passengers and cargo). A specific country name has been used in the feature class, for example, (USGE) United States government establishment (Gloss: facility operated by the United States Government in Panama). For the many feature codes glosses are missing. For instance, there are no glosses available for feature code such as PPLA2, PPLA3, PPLA4, PPLG, PPLX, LGNX etc. There is no rigorous or formal terminology management in GeoNames. The GeoNames does not provide temporal information. So it is impossible to understand that entity still exists or is destroyed. For example World trade Center (latitude 40.4241 longitudes 74.044) was constructed in 1968 and destroyed in 2001. From 2001 its called ground zero or World trade center site for nearly 13 years. It reopened for business on 3 November 2014. GeoNames shows the Freedom tower (nick name of the World Trade center) and World trade center site under the tower and building respectively with the same latitude and longitude information.

GeoWordnet [Giunchiglia et al., 2010] has built the connection between the WordNet (purely lexical resource) with the GeoNames. It constitutes a first attempt to approach the terminology problem of GeoNames. GeoWordnet is a geospatial resource, which adds the semantic relation with the GeoNames feature class and groups them according to WordNet hierarchy. Their main objective was to achieve semantic interoperability. In 357 cases, GeoWordnet introduced a new concept, which is not available in exiting WordNet and positioned them in the right place of in the hierarchy. For example the term geo-political entity is introduced and has been connected through hypernym relation to physical objects. GeoWordNet has no provision to represent time. It is not built for any domain specific

application.

OpenStreetMap (OSM) [Haklay and Weber, 2008] is a rich source of Geo-spatial information for both free use and commercial use. All geographical entities are entered in the database of OSM as points (nodes) that have spatial attribute such as the longitude and longitude coordinate [Haklay and Weber, 2008]. The maps use several tags to represent physical features on the ground (e.g., roads or buildings). Core part of the OSM is the tagging schema, which has been developed into the taxonomy of real-world features classes and objects. It provides gloss for the tags used in their map. In a few cases, no specific definition is available for the tags. Sometime it used as an acronym, for instance atm as an ATM or cash point. There is a need to define vocabulary in formal language, which helps the machines to interact with each other.

OSMonto [Codescu et al., 2011] is an ontology key (k) and value (v) used in OpenStreetMap tags. In the ontology they have not corrected the conceptual mistake in the design of OSMs tags, rather they introduced their own syntactic tags, for instance *v_smoking_k_no* and added the prefix *k_* and *v_* for all key and value respectively [Codescu et al., 2011]. OSMonto use the web ontology language (OWL) as representation languages. OWL does not allow same name for different nodes. For that reason some ad-hoc tags were used in the ontology, for instance *v-no*, without providing proper justification and the meaning of their introduced tags. The main application domain of this ontology is on spatially located activities and is helpful for our daily activities such as finding restaurant, nearest ATM or Post-office on the way of your home etc.

FAO Geospatial ontology [Iglesias-Sucasas et al., 2013] is the core part of an information retrieval system developed by the FAO (Food and Agri-

cultural Organization of the United Nation) for access to the FAOs thematic country-based information. The ontology was developed in OWL and included relationship among geopolitical entities (e.g. countries, territories, region) [Iglesias-Sucasas et al., 2013]. It does not provide any terminology support. The main objectives of the FAO Geopolitical ontology are to provide the updated geopolitical information and to improve information management. Thus, it does not support any domain specific scenario mentioned in our introductory part.

3.4.3 Applications

T-YAGO [Wang et al., 2010] YAGO (Yet Another Great Ontology) is a knowledge base developed at the Max Planck Institute. It is automatically extracted from Wikipedia and WordNet. Timely YAGO (T-YAGO) is a version of YAGO, where they enrich their ontology by adding temporal information. T-YAGO gives emphasis on sports personalities, entertainment, and political personalities to support journalist or media analysts. It provides temporal information of the persons biography such as date of birth; the time period in which one player played with the particular sport club etc. Award received or political position held by a politician during a specific period. For instance, David Beckham has played for Real Madrid since 2003 until 2007. Barak Obama *wasBornOnDate* 04-08-1961. David Beckham has won the UEFA Club player of the Year in 1999. The Resources description framework (RDF) data model is used in T-YAGO. RDF has a property that it is a binary relation (link to individual or an individual and a value) [Noy et al., 2006]. Facts might contain only binary relation but temporal fact have more than two arguments [Wang et al., 2010]. For supporting temporal facts in YAGO, the n-ary relation, that allows using relation to link an instance to more than one instance or value, is decomposed. A fact identifier has been assigned to the primary fact that,

build the relationship between identifier and remaining argument [Wang et al., 2010].

YAGO2 [Hoffart et al., 2013] is the new version of YAGO, which integrated GeoNames in their knowledge base. Its knowledge base contains more than 350,000 classes, 10 million entities and 120 million facts and combines the clean taxonomy of WordNet with the richness of the Wikipedia category system. Name matching and geographical coordinate matching [Liu and Yoshioka, 2011] is used in YAGO2 for integrating GeoNames and Wikipedia. YAGO did some cleaning in the taxonomy of WordNet but the clean taxonomy of WordNet used in YAGO2 still has several limitations such as semantic mismatch for instance, *New_Roman_Times* which is an album by musical group Camper Van Beethoven, released October 12, 2004 is linked with *Time_New_Roman* is a serif typeface commissioned by the British newspaper The Times in 1931. Wrong classification of Geo-spatial entity for instance *Chandhi Chowk*, (moonlit square) which is one of the oldest and busiest markets in Old Delhi, India is classified under WordNet show and social event (e.g. movies show) as it matches syntactically with one of the Bollywood movies named *Chandhi Chowk to Chaina*.

YAGO2 does not provide temporal information for the capital city of the country as many countries have two national capitals. For instance, Bonn and Berlin are the national capitals of Germany, Kolkata and New Delhi are the national capitals of India. Florence and Rome are the national capital of Italy. We found that out 6 out of 62 national capitals, in *wordnet_national_capital_108691669* dataset have wrong information. This means that at least 9.6% information is wrong. This is mainly because of the use of WordNet as a knowledge base. In the WordNet, they do not distinguish between capital (“*a seat of government*”) and national capital (“*the capital city of a nation*”). Coming to World trade center example,

YAGO2 have two, *World Trade Center* (latitude 40.7 longitudes 74.0) and *One_World_Trade_Center* (latitude 40.7 longitudes 74.0) with the same coordinate information. This type of mistake is not acceptable.

YAGO2 attaches a temporal dimension and a spatial dimension to many of its facts and entities but it does not build for any domain specific information. It does not provide temporal information for entities like capital, building and archeological site needed for answering queries of our mentioned scenario. The accuracy of YAGO2 has been manually evaluated; proving a confirmed accuracy of 95%. But their confirmed accuracy of 95% is really questionable. Out of the 120 million facts, only 3790 facts are evaluated manually and assuring 95.02% accuracy. This means that only 0.0032% facts have a chance of being accurate. In *subClassOf* relation, only 41 facts have been checked out of 4,58,979 facts and claimed to be 100% correct (95.72% weight average). But we can argue that, in case of a large diverse population (e.g. 458979) where the sample size is less than 0.83% (e.g. 384), we cannot assure that the facts are correct or not.

Pleiades [Center et al., 2014] is a map based application for visualizing ancient site supported by the National Endowment for the Humanities, the institute for the study of the Ancient World, New York University, Ancient World Mapping Center and The University of North Carolina at Chapel Hill. It is a community- built gazetteer and graph of ancient 34,764 places. The dataset is available in various formats such as Turtle, RDF, KML and CSV file. It is a rich and authentic information source of the ancient site. But it does not provide terminological support and have to provision for accommodating temporal information.

Table 3.1: Pilot Geo data sources

Data Source	Format	Description
www.openstreetmap.org	XML	Physical features on Earth surface (global)
dati.trentino.it	CSV XML JSON geojson shp	Local province, providing the most accurate and up-to-date geographic data (local)
www.europeandataportal.eu	CSV XML JSON geojson	European commission, providing authoritative data (global)
www.protectedplanet.net	CSV	UNEP maintaining and providing latest information regarding world protected site (e.g. national forest, world heritage site) (global)

3.5 Informal Modeling Phase

3.5.1 Pilot Reference Datasets

In the early days of this work, we investigated various sets of (open) geospatial data (see Table 3.1) in an effort to better understand the challenges of their integration using ontological approaches. The investigation involved listing the datasets that we intended to integrate, their formats, terms used in the datasets to denote the geographical features and matching our needs, and consulting various geospatial resources.

3.5.2 ExER Model

A partial view of Extended Entity Relationship (ExER) model shown in Figure 3.2. This view depicting how location is interconnected with other

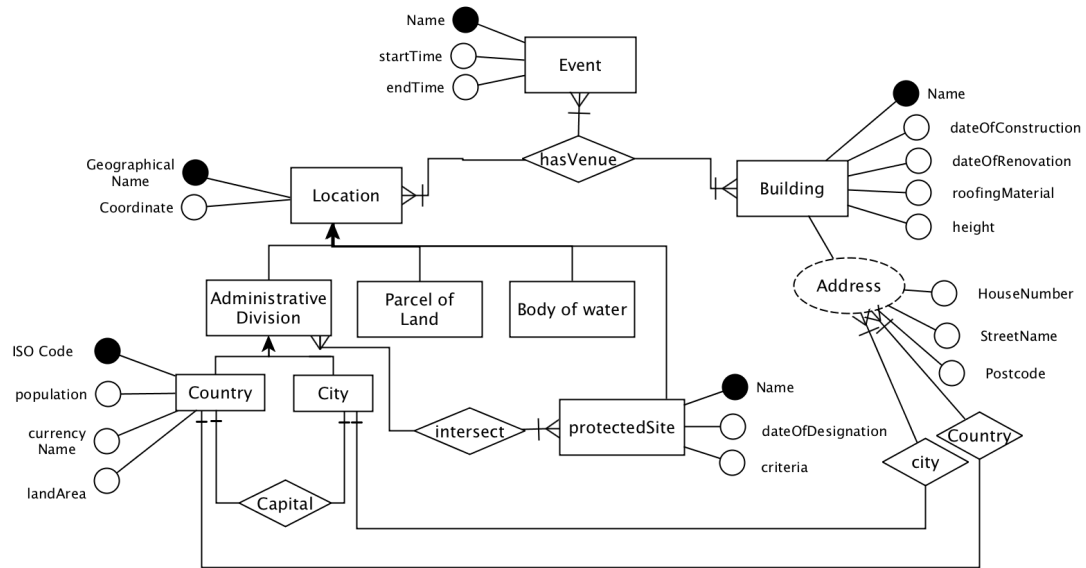


Figure 3.2: ExER Model for Geospatial

main entity types; Event and Building. An Event can organize in an open air field that is venue is a certain location or it can be host inside a building. Building has address which refer to a location where it is situated. Administrative division such as country, city, body of water such as river, lake are the subclass of a location. Even administrative division like city, country are the component of a structure attribute like the address. Building can have temporal attribute like date of construction, date of renovation and height.

3.6 Geo eTypes

Each user will take a different view of the world this is largely driven by their application. Increasingly these users need to exchange and share information about the same real world entity. This can be achieved at least for spatial objects where the spatial characteristics are related to topographic objects by one of three general approaches:

- **Simple overlay:** In the overlay method each user-defined geography may be based on a different base map. Even where it has been digitised from the same topographic map, co-ordinate differences may be difficult to avoid and therefore gaps and overlaps may often exist (these are not present in the real world features).

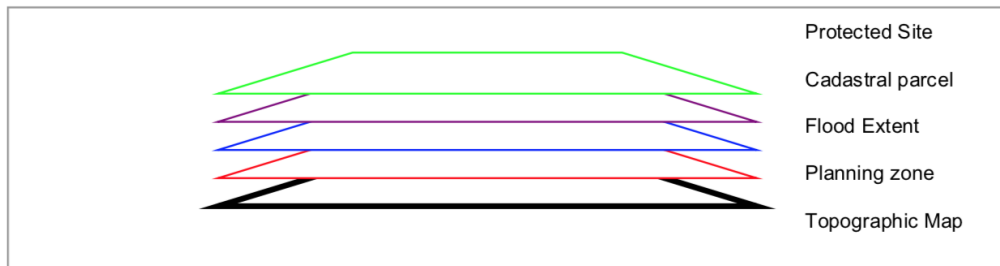


Figure 3.3: Simple Overlay

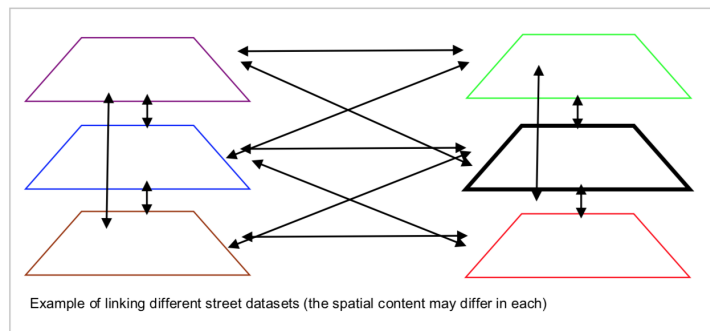


Figure 3.4: Many-to-many linking

- **Many to many linking :** Many to many linking using external identifiers (e.g. of various views of a highway maintained by different organizations) establishes an explicit relationship. Nevertheless the disadvantages of the overlay method remain and this is compounded by the need to maintain as many cross references as there are geographic relationships. These spatial objects are rarely coterminous, maintenance is disjoint and hence data sharing is very often very inefficient and ineffective. In general, associations between spatial objects

described by different data specifications shall be kept to a minimum as a large number of associations will make it more difficult to maintain consistent data and to reduce the effects of updates on other data sets, etc.

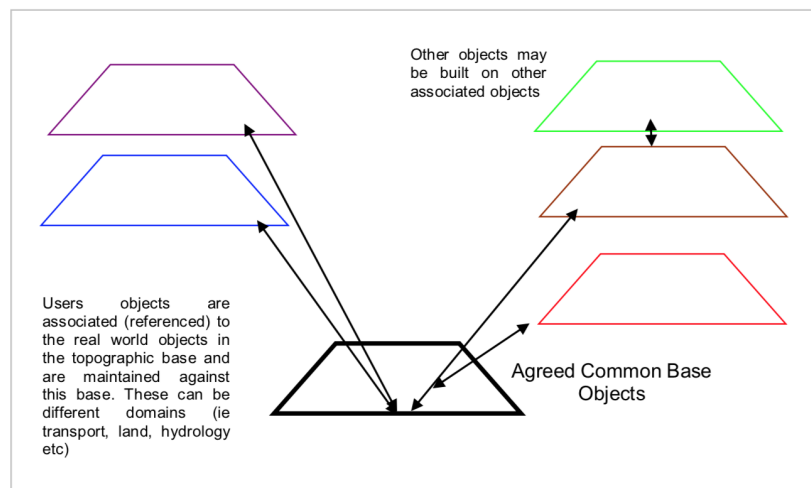


Figure 3.5: Referencing a common base

- **Referencing a common base:** In object referencing it is assumed that there is a commonly agreed and well defined base of spatial objects that others can associate their own information with that base. Users who build on these spatial objects will inherit the links, continuity and integrity provided by the common base. The approach supports the key INSPIRE principles of data sharing and information reuse across distributed spatial data sets. The different steps for achieving data integrity by object referencing are the following :
 - at national level, find an agreement on which data forms part of the national reference data
 - each public organization will build their data on the reference data, using the object referencing methods described below
 - each public organization will publish their data according to the

user's requirements

To deal with *diversity in the schema*, we choose INSPIRE [Directive, 2007] as a reference standard. INSPIRE, the European directive for spatial information in Europe. It is an international accepted standard developed by the experts and based on Open Geospatial Consortium (OGC) open standards. It has 34 data themes described under 3 annexes, such as hydrography, protected sites, administrative units and so on, to enhance interoperability of spatial datasets among the member states. Following the directive, we derive our *Geo eType*.

Geo eType is the set of all *eTypes* of the *Geo entity* (i.e., *a physical object (here physical object means a tangible and visible physical entity), which has an existence in our planet Earth and occupies a certain geometric area which we represent as a point, line, and polygon, and which maintains their identity through time*). Geographical regions separated for administrative purposes (e.g. country, province) or classified based on the Earth's vegetation pattern (e.g. Alpine region, Savanna region, Tundra region) as well as man-made objects such as buildings or constructions (having permanent position on the Earth's surface) are considered as a *Geo entity*. We provide in the rest of the section a description of our foundation, based on the definition (1), (2) of *eType* and details analyzing classes, attributes, datatypes, and the overall model.

Here, we provided details attributes and relation for main geospatial eTypes. They are Location, Administrative Division, Building, Body of water etc.

Table 3.2: Location eType

Name	Description	DataType
Identifier	a symbol that establishes the identity of the one bearing it	STRING
Geographical name	a name by which a geographical location is known.	[] NLSTRING
Description	The description of the entity	[] NLSTRING
Coordinate	a number that identifies a position relative to an axis	GEOMETRY

Table 3.3: Administrative district eType

Name	Description	DataType
Geographical name	a name by which a geographical location is known.	[] NLSTRING
Country code	country code as per ISO standard	STRING
Surface	the extended two-dimensional outer boundary of a three-dimensional object	GEOMETRY
National level	number according to National level. (e.g. 1-5)	INTEGER
Area	the extent of a 2-dimensional surface enclosed within a boundary	FLOAT
Population	the number of inhabitants in a given place	INTEGER
Residence of authority	government building from where government function	Building

3.7 Schema Level

3.7.1 Defined Classes

Classes are the core component of the *eType*. *Geo eTypes* has two root classes. One is artifact, which is the parent class of all man-made construction (e.g. building, structure). The second one is geographical location (or simply location). Geographical location is divided into nine core classes such as *administrative division*, *bio-geographical region*, *body of water*, *geological formation*, *habitat*, *natural risk zone*, *parcel of land*, *protected site*, and *transportation area*. Core classes are aligned with the

Table 3.4: Building eType

Name	Description	DataType
Name	a name by which a entity is known.	[] NLSTRING
Date of construction	the date on which construction begins	DATE
Date of renovation	the date on which building renovated	[] DATE
Height	the vertical dimension of extension	FLOAT
Address	written directions for finding some location	Address
Roofing material	building material used in constructing roofs	Concept
Current use	current using purpose (e.g. governmental, commercial, educational)	Concept

top-level of DOLCE ontology [Guarino, 1998, GIUNCHIGLIA and FUMAGALLI, 2016]. The main classes of *Geo eTypes* shown in Figure 3.6. Ontology editing tool *Protégé* (Version 5.0)⁴ has been used for creating the *Geo eTypes* ontology⁵.

⁴<http://protege.stanford.edu>

⁵The complete ontology is available at <http://webprotege.stanford.edu/#Edit:projectId=f5f2c273-9f6a-4d4b-af7b-451166f58286>

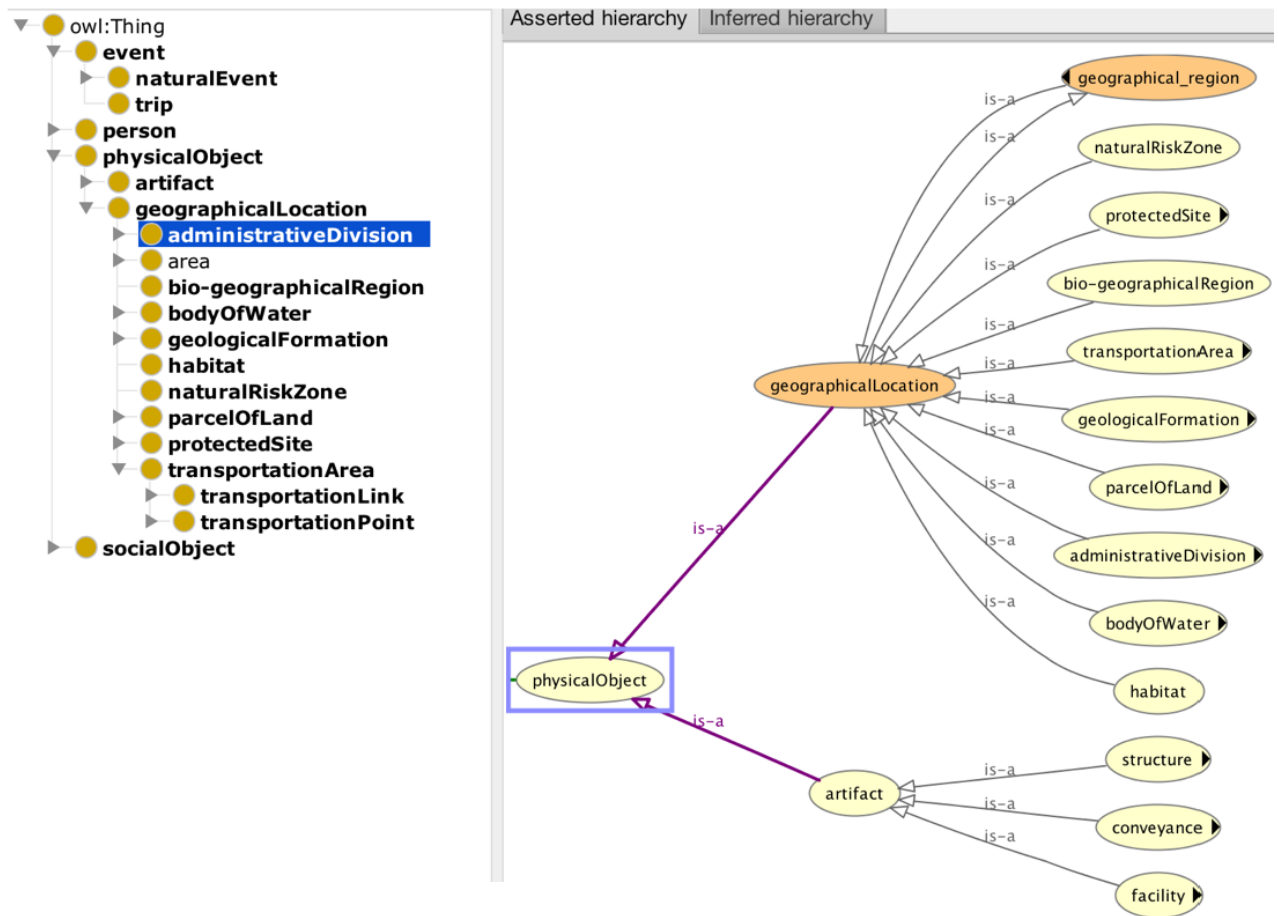


Figure 3.6: Main classes of Geo eTypes

Class relationship for geospatial domain shown in Figure 3.7.

3.7.2 Alignment with Upper Ontology

Geospatial domain has two root class, both are connected with top-class physical object. Other interdependence class like trip, natural event has root class event. And for the authority responsible maintenance and administered some area or building are rooted under social object class. For definitions of concepts mentioned here see Section 3.1.1 and 3.3.1.

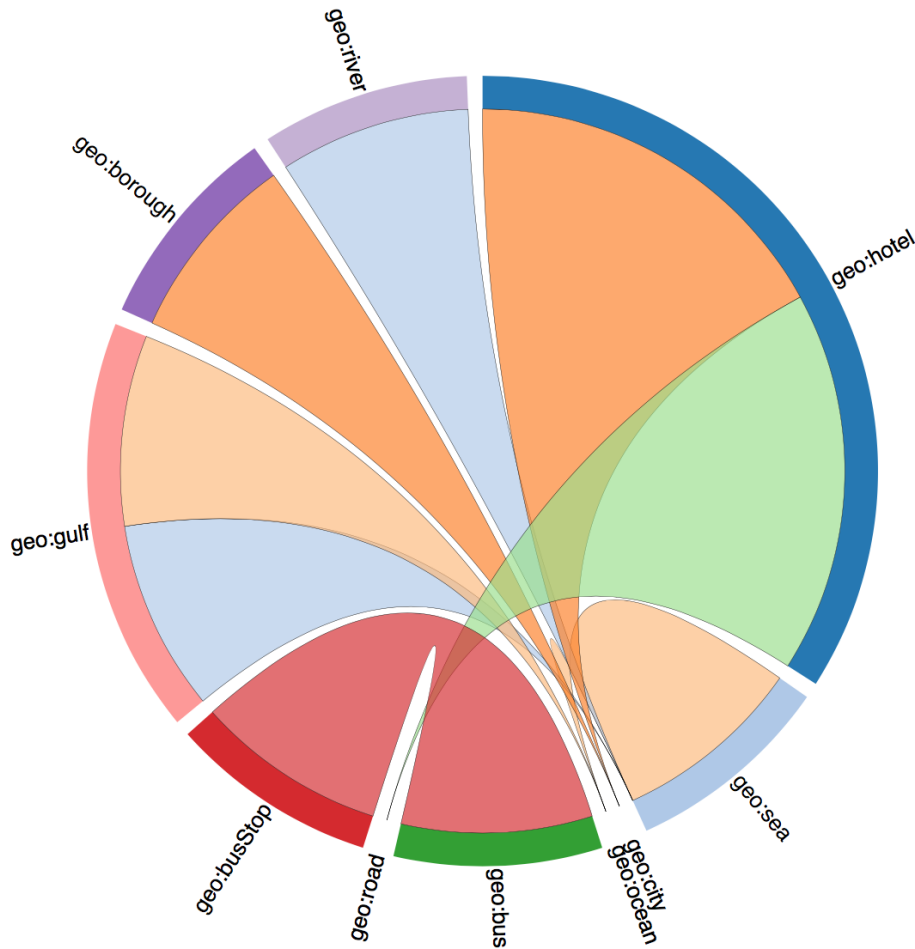


Figure 3.7: Class Relationship Geospatial

3.7.3 Defined Attributes

Spatial Attribute: Spatial types are represented as points on a planar, or flat-earth, surface. An example would be (7,6) where the first number represents that point's position on the horizontal (x) axis and the second number represents the point's position on the vertical (y) axis. Geometry data types are represented as latitudinal and longitudinal degrees for a point, as on Earth or other earth-like surfaces; as polygon for a geographical area and as poly line in case of a linear spatial object.

Altitude and Elevation also consider as a spatial attribute as it need to

Table 3.5: Spatial Attribute

Attributes	Domain [eType]	Range [Data Type]
	Point location (city, bus stop)	
Latitude	Any structure which has identified or permanent/semi-permanent position on earth surface (Building, Statue)	float
	Point location (city, bus stop)	
Longitude	Any structure which has identified or permanent/semi-permanent position on earth surface (Building, Statue)	float
Altitude		LengthDType
Elevation	Land on Earth surface (e.g. Mountain, Hill, Summit)	LengthDType

represent in respect with one particular spatial point location. Table 3.5 shows example of different spatial attribute and their respective domain and range. Where as domain indicates the class or entity types where this type of attribute is applicable and range showing the data types. As we can see latitude and longitude can be measure using float type and for altitude and elevation can be measure using any unit for measuring length i.e. meter or feet. We call them as *LengthDType* in short. This type of categorization of the attribute is really helpful for the entity visualization (more information regarding this available in Chapter 7).

Temporal Attribute: The time-related attribute which contains temporal information of an entity and can be represented using date, DateTime are categories under temporal attribute. Table 3.6 shows some example of temporal attributes which we used in our model.

Table 3.6: Temporal Attribute

Attributes	Domain [Etype]	Range [Data Type]
Date of observation	Body of water	Date
Date of inscription	Protected site	Date
Date of registration	Parcel of land	Date
Date of ceasing	Parcel of land	
Date of construction	Building	Date
Date of renovation	Building	Date
Date of demolition	Building	Date
Hours of service	Governmental service	TimeData
OpeningHours	Building [provides service]	String
Period of validity	Hazard area	Date [begin Date + end Date]
Historical time	Protected site	Date [begin Date + end Date]
Holiday	Building [provides service]	Date [begin Date + end Date]

3.8 Language Level

We imported 400 concepts from United States Geological Survey(USGS)⁶ and United Nations International Strategy for Disaster Reduction (UNISDR)⁷, apart from included all GeoWordNet [Giunchiglia et al., 2010] concepts. Then we arrange them according to DERA facet [Giunchiglia and Dutta, 2011]. Facet development is process where we organized term based on

⁶<https://www.usgs.gov>

⁷<http://www.unisdr.org>

DERA category: entity, relation and attribute. For example, partial view of DERA facets for geospatial domain shown in Figure 3.8. We imported 400 concepts

<u>ENTITY</u>	<u>RELATION</u>	<u>ATTRIBUTE</u>
Artifact	locatedIn	name
(is-a) Structure	hasVenue	latitude
(is-a)Building	residenceOfAuthority	longitude
(is-a)Factory	administerBy	elevation
(is-a) Conveyance	mouth	surfaceArea
(is-a)Vehicle	topological	population
GeographicalLocation	(is-a) adjacent	landUse
(is-a) AdministrativeDivision	(is-a) contains	(value-of) industrial
(is-a) BodyOfWater	(is-a) inside	(value-of) residential
(is-a) BiogeographicalRegion	(is-a) equal	(value-of) agricultural
(is-a) ParcelOfLand		

Figure 3.8: Example of Facets

3.9 Evaluation

Our evaluation criteria are based in the ability of *Geo eTypes* to cover classes and attributes from a large set of dataset. We run our study on four datasets on which we did the investigation These datasets are: OpenStreetMap (OSM) dataset, ProtectedPlanet (ProPlanet) dataset, European data protal (EDP) and OpenDataTrentino (ODT) (available in Italian).

The quality of our proposed *Geo eTypes* model is evaluated based on four measurements: (a) completeness, (b) minimality, (c) understandability and (d) expressiveness, as proposed by [Akoka et al., 2007]. Let us explain them below.

Completeness. Completeness of our model is checked based on two phases: class mapping and attribute mapping. During class mapping we

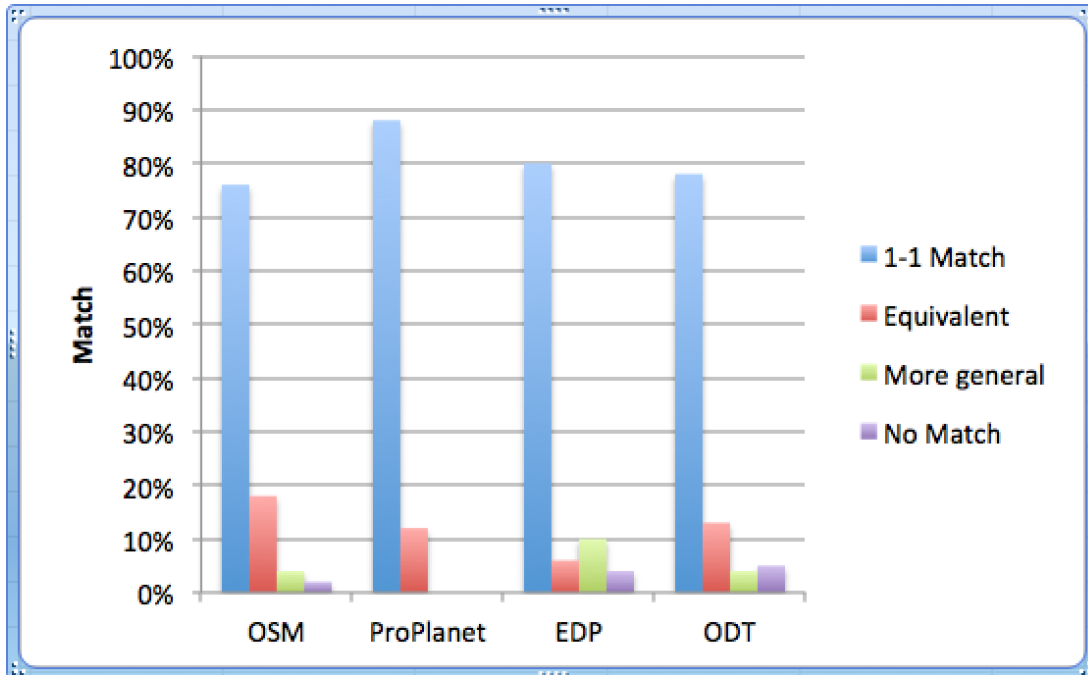


Figure 3.9: Statistic

grouped similar types of entities into one *Etype*. For example, OSM entities with the tag “airways” and “aerodrome” become *airport* and tag (*OSM_key=boundary*, *OSM_value=administrative*; *OSM_key=boundary-type*, *OSM_value=national*) grouped under *country*. We checked geographical features with our 770 *concepts* those we defined in the *Geo eTypes voc* and statistics of our evaluation result are shown in Figure 3.9. *1-1 match* means when the target concept is found in defined vocabulary (country-country), *equivalent match* refer to when a similar concept (zippostal code) is in the vocabulary, and *more general match* means when the immediate superordinate concept available than the concept is available in the input schema (pastry shop-bakery shop).

Some examples of equivalent (i.e., meaning is same) attribute names are *year of designation* and *date of inscription*, *addr:street* and *thoroughfare name* and *CAP* and *postal description* so on and so forth.

Minimality. Notice that with 770 *concepts* we are able to capture all geographical features of OSM (i.e., 1112 features) and features from other datasets.

Expressiveness and Understandability. The model provides natural language description for all 770 concepts used in the model and it complies with the requirement of *concept expressiveness* [Akoka et al., 2007]. Schema of our model expresses on entity-relation model, made more expressive and understandable for common people.

In the second part of the evaluation, we perform query evaluation but here we are providing only one example to show that our model can support analytical query. list all the eating establishment which has cuisine rating higher than 4 along with their service quality.

```

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rest: <http://www.semanticweb.org/subhashis/
ontologies/2016/9/restaurants#>
SELECT DISTINCT ?eatingEstablishment ?ServiceRating ?CousineRating
WHERE { ?eatingEstablishment rdf:type ?type.
       ?type rdfs:subClassOf* rest:eatingEstablishment .
       ?eatingEstablishment rest:serviceRating ?ServiceRating.
       ?eatingEstablishment rest:cuisineRating ?CousineRating
FILTER (?CousineRating > 4).
}
Order By ?eatingEstablishment

```

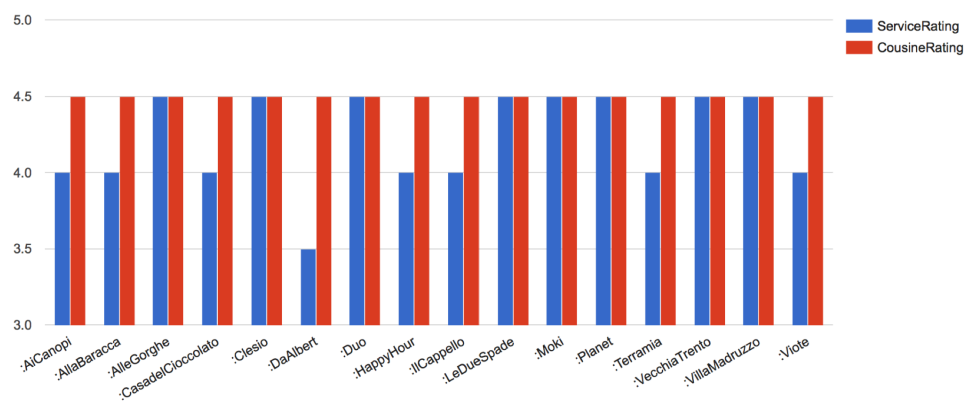


Figure 3.10: Comparison of service quality and cuisine quality of restaurant in Trento

Chapter 4

Smart Transportation Domain

“Wherever you go, go with all
your heart.” —*Confucius*

Transport or transportation is the movement of people, animals and goods from one location to another. Modes of transport include air, rail, road, water, cable, pipeline and space. The field can be divided into infrastructure, vehicles and operations. Transport infrastructure consists of the fixed installations or structure including roads, railways, airways, waterways, canals and pipelines and terminals such as airports, railway stations, bus stations, warehouses, trucking terminals, refueling depots (including fueling docks and fuel stations) and seaports. Terminals may be used both for interchange of passengers and cargo and for maintenance. Vehicles play role of a agent to connect different transportation points or nodes, traveling on these networks may include automobiles, bicycles, buses, trains, trucks, people, helicopters, watercraft, spacecraft and aircraft.

4.1 Domain Description

Wikipedia defines transport network, or transportation network as “a realization of a spatial network, describing a structure which permits either

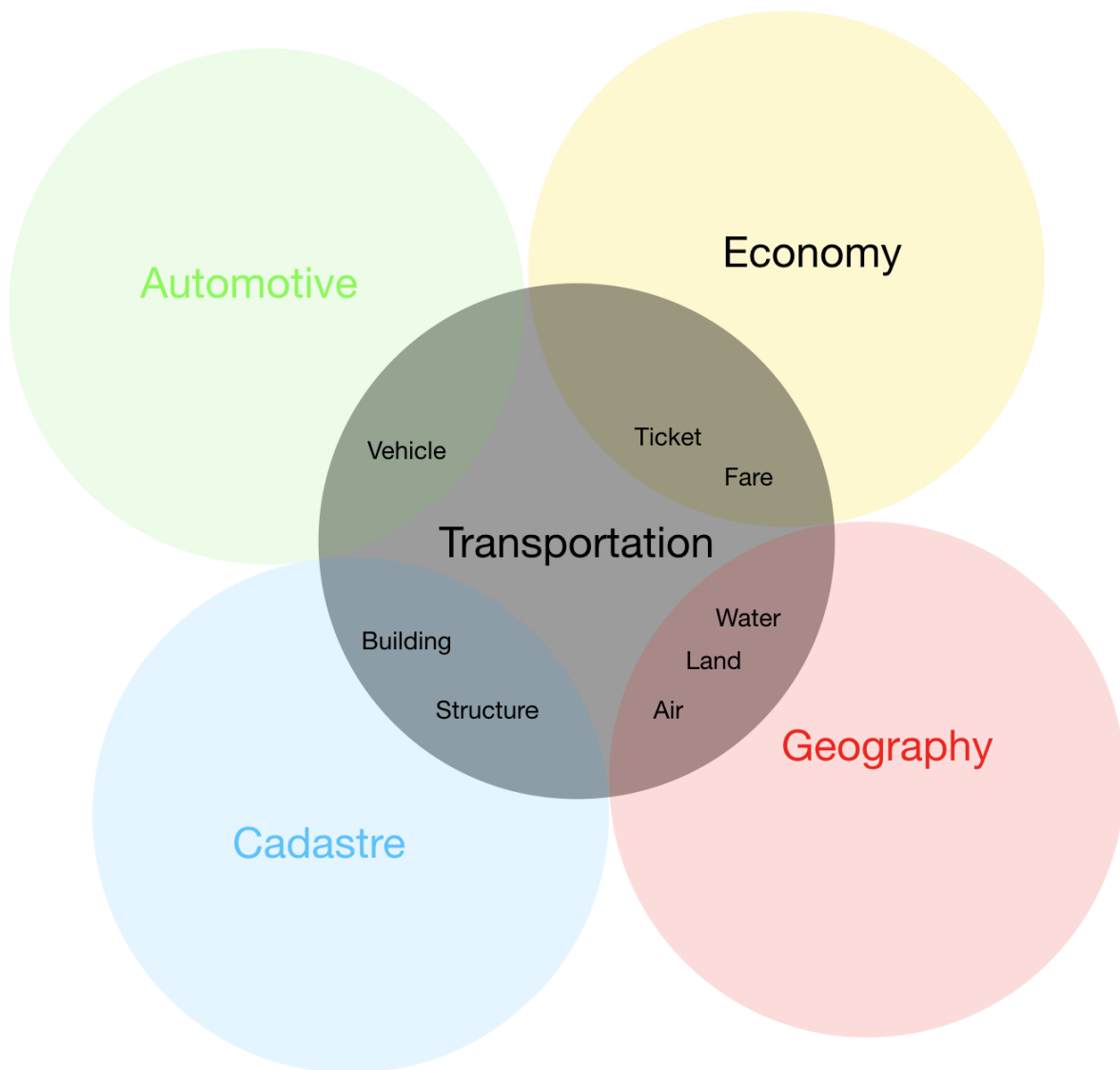


Figure 4.1: Transportation Domain

vehicular movement or flow of some commodity. Examples are network of roads and streets, railways, pipes, aqueducts, and power lines”. This definition clearly shows two distinct components in a transport network: spatial network and vehicular. However, in a real-life situation, there are many components inside a spatial network such as locations, organizations, structures which are inter-connected with vehicles and events. Moreover, most of the existing model treat those components only as a location by using some tags or label as they can be represented on a map. This creates problems while integrating different Transportation datasets (because of different natural language labeling). Figure 4.1 illustrated how transportation domain intersects with other domains using the set diagram.

Transportation is one of the most widely discussed theme in Open Data [Wendy Carrara, 2016]. A statistical report from the European Data Portal¹ suggests that it is the most downloaded category. Furthermore, among different transportation categories, public transportation data has the highest impact on a citizen’s life. As the data majorly deals with the crucial information such as timetable, frequency and other services associated with the means of transportation. The report from European Commission shows that the frequency of people commuting every year using different public transport has increased significantly². Furthermore, in recent days, the nature of travel pattern has become more complex.

4.1.1 Problems

Publicly available data contains various issues. The data is available in the various format, follow different standards, and language creating complexity while integrating them. This complexity in data also arises several issues: (1) Various public transportation services providing different trans-

¹<http://www.europeandataportal.eu>

²<https://goo.gl/ZToIpm>

portation means, creating confusion among common people to choose best travel option. (2) Currently, existing solutions mostly handle a specific mean of transportation i.e. the user has to check multiple applications and then analysis which combination best suits her. (3) Interlinked multidimensional complexity due to overlapping of temporal (time table) and spatial data (stops and stations). (4) The geographic information related to transportation is not interlinked with other services such as gas station or hotel. These issues create obstacle while making complex queries.

4.1.2 Purpose and Objectives

Many different ontology based transportation models are available proposing different solutions. However, these models mostly fail to differentiate between various transport components. For example, buildings, road links, and stops are entirely treated as a location. Moreover, these models use different standards as well as local terminologies to classify and model the data.

We propose Ontology-aware transport model (ATOM) to tackle the terminological diversity and to accommodate and link different transportation datasets. ATOM underpins ontological principle, which facilitates answering complex queries related to different transportation means and facilities. The model also supports processing query like: services such as hotels and restaurants near to a transportation point. We follow INSPIRE (Infrastructure for Spatial Information in the European community) directive [Directive, 2007] to model ATOM.

4.2 Reference Scenarios

For the last decade the Municipality of Trento has been fighting inner city traffic. Situated in a mountain valley, the city has limited options

of expanding its road infrastructure and hence needs actively to discourage excessive car traffic in its city centre, a policy also enforced by regional law. While recent measures such as the creation of a limited traffic zone, the deployment of bike sharing services, or the establishment of paid parking zones with variable fees have had positive effects on traffic, the municipality has only limited means of quantifying these improvements and of understanding the underlying reasons. We present a definition and an overview of the definition and relevance of modal split to address the issue of traffic in European cities and the types of services that the Municipality aims to implement as part of the use case.

Parking is also known for having a profound impact on city traffic and for being a source of pollution. As such, efficient parking policies are crucial when dealing with both these issues. However, the Municipality has little knowledge about usage of parking spots around the city, mainly limited to off-street, underground parking. A comprehensive analysis of parking availability would help the Municipality setting priorities for future policies. Furthermore, the Municipality wishes to set up a do and don't strategy; while collecting data through sensor network and it wants to provide citizens with valuable services in exchange of their precious contribution. An assessment of the needs of the citizens of Trento has been carried out mainly by getting input through an ideas competition; results not only show an increased awareness of citizens in terms of traffic reduction and alternative mobility but also a need to be informed about mobility in Trento. The QROWD project brings solutions to these problems through the combination and analysis of big data from the Municipality's database, participatory sensor data from the mobile devices of citizens, and lightweight electronic surveys on mobile devices. The modal split can therefore be computed yearly or even monthly as opposed to every ten years, at a fraction of the cost, with higher precision due to a continuous,

comparable computation, and with finer granularity in terms of geographic coverage as it involves commuters other than resident citizens. Finally, citizens of Trento will benefit from the QROWD project through an improved mobility experience in their daily life. Citizens are also motivated to chose sustainability options by getting free access to others, personalized services offered by the QROWD project.

4.3 Generalized Queries

We defined and validated a set of generalized queries to check the effectiveness of the model. The generalized queries are:

- Q1 Is the stop disabled friendly?
- Q2 Whether bicycle is allowed in the trip?
- Q3 What are the nearest facilities surrounding of a transportation stop?
- Q4 How to buy a ticket in the transportation means?
- Q5 Whether pets are allowed in public transport?

4.3.1 Identify the Domain Concepts

A transport network, or transportation network is a realization of a spatial network, describing a structure which permits either vehicular movement or flow of some commodity. Examples are network of roads and streets, railways, pipes, aqueducts, and power lines¹. Every component of the transportation network is not belongs to location. Few elements are artifact such as Bus, Train and Taxi etc. and some have been considered as a location. For avoiding any ambiguity, in our work we consider the

transportation area limited to the location from where we can avail transportation service. It consists of two elements, namely transportation point (bus stop) and transportation link (such as railway lines, road etc.)

Transport point: A point spatial object - which is not a node - that represents the position of an element of a transport network. For example, bus stop, buoy, beacon etc.

Transportation link: A linear spatial object that describes the geometry and connectivity of a transport network between two points in the network. For example, railway line, road.

Transportation node: A point spatial object which is used for connectivity. For example, bus station, railway station, airport

Other concepts related to transportation are:

Airport/Heliport: a defined area on land or water (including any buildings, installations and equipment) intended to be used either wholly or in part for the arrival, departure and surface movement of aircraft/helicopters[Directive, 2007].

Apron Area: a defined area, on a land aerodrome/heliport, intended to accommodate aircraft/helicopters for purposes of loading and unloading passengers, mail or cargo, and for fueling, parking or maintenance [Directive, 2007].

Bridge: a man-made structure spanning and providing passage over a body of water, depression, or other obstacles[Directive, 2007].

Deep water route: a route in a designated area within defined limits which has been accurately surveyed for clearance of sea bottom and submerged obstacles to a minimum indicated depth of water[Directive, 2007].

Facility: equipment or service that provides a specific convenience or service to passenger [STANDARDIZATION, 2011]. For example Ticket machines, elevator, mechanical stair, toilet, portage, left luggage, etc.

Roundabout: a road junction at which traffic streams circularly around

a central island.

4.3.2 Query Collection

We collected all necessary queries from the municipality of Trento, as they already have queries log where they stored all type of transportation data.

4.4 State of the Art

GTFS: The General Transit Feed Specification (GTFS), also known as GTFS static or static transit to differentiate it from the GTFS realtime extension, defines a common format for public transportation schedules and associated geographic information[Google, 2006]. It is use as a *De facto standard* on Transportation. GTFS “feeds” let public transit agencies publish their transit data and developers write applications that consume that data in an interoperable way.

Service Interface for Real Time Information (SIRI): The Service Interface for Real Time Information (SIRI) specifies a European interface standard for exchanging information about the planned, current or projected performance of real-time public transport operations between different computer systems. SIRI comprises a carefully modularised set of discrete functional services for operating public transport information systems. SIRI aims to incorporate of the best of various national and proprietary standards from across Europe and delivers these using a modern XML schema and TransModel terminology and modeling concepts. All SIRI services are provided over a standardised Communications layer, based on a Web Services Architecture.

FIWARE: The FIWARE Community is an independent open community that makes and shares open source technology for smart solutions to build an open sustainable ecosystem around public, royalty-free and

implementation-driven software platform standards that will ease the development of new smart applications in multiple sectors including, but not limited, to smart cities. FIWARE Data Models have been harmonized as per our experience working with different services in Smart Cities and other domains. A further requirement was that the data model could accommodate the static data provided by the Municipality, and available via the Open Data Trentino platform³. This is still a work in progress, due to the fact that some new dataset will be available in March 2018, e.g., real time navigation of buses, camera feed from lampposts, and bike sharing dock stations.

4.5 Informal Modeling Phase

4.5.1 Pilot Reference Datasets

In the early days of this work, we investigated various sets of (open) geospatial data (see Table 3.1) in an effort to better understand the challenges of their integration using ontological approaches. Our objective was to scrutinize mainly dataset from Open Data Trentino⁴ which has total 13 datasets. And we also collected some other datasets directly from the municipality of Trento.

4.5.2 ExER Model

The trip entity has attribute tripId, headsign, and others. A trip always follows some route. The operator decides fare for travelling certain route. During travel, you can pay fare using different payment methods such as in the bus, online, or advance at the ticket office. Vehicle entity follows certain route within a trip. Here, we are considering only public vehicle.

³<http://dati.trentino.it/>

⁴<http://dati.trentino.it/group/mobilita>

Organization (government or agency) operate all vehicle and also maintenance road, bus stop, and other transportation areas.

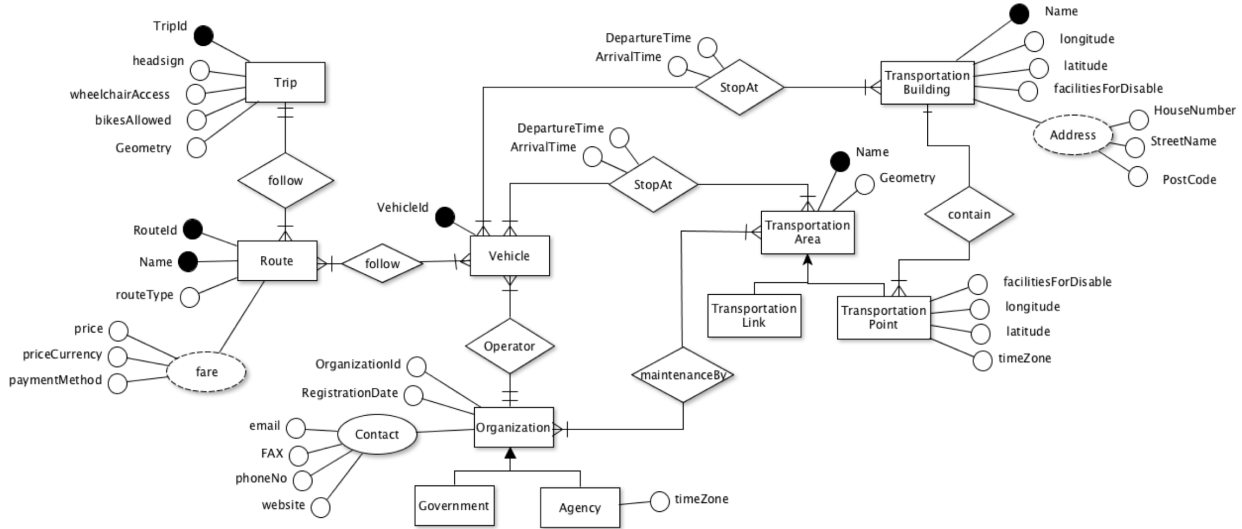


Figure 4.2: ExER Model for Transportation

4.6 Transportation eTypes

Here we describe main transportation eTypes required under QRAOWD project. The Trip eType (see Table 4.1) represents a trip, which is defined as a movement from a geographical point A to a geographical point B. It is bounded either by the citizen reaching destination, or by changing his mode of transport.

The Parking eType (see Table 4.2) represent all the types of parking spots available in a city, i.e., offstreet parking sites with explicit entries and exits, on street, free entry (but might be metered) parking zone which contains at least one ore more adjacent parking spots, and bikesharing docking station.

Table 4.1: Trip eType

Name	Description	DataType
Name	a name by which a entity is known.	[] NLSTRING
Origin	the place where something begin	Location
Destination	the place designated as the end	Location
Path	an established line of travel or access	GEOMETRY
Payment method	The way that a buyer chooses to compensate the seller of a good or service that is also acceptable to the seller.	Concept
means of transportation	conveyance for the travel	Vehicle
Start Date	Start date of the event	DATE
End Date	End date of the event	DATE

4.7 Schema Level

Similarly, INSPIRE defines network theme as road, rail and water transport and includes a different set of infrastructure. They have five major sub-themes as road, rail, water, air and cableways. Their specification also considers a set of non-geographic data such as asset condition and traffic flow reports. Our use case is more focused on user preferences during travel management.

In addition, General Transit Feed Specification (GTFS) also provides transport specification that defines a common format for public transportation schedules and associated geographic information. This specification is only focused on agency, routes, trips, stops, stop times and calendar. So other facilities such as disabled friendly, allowed entry for a pet or the way to buy tickets is completely ignored.

Our model combines the specification from both the INSPIRE and the GTFS to provide an integrated model and to support complex queries. The focus of the ATOM is to improve information retrieval by capturing

Table 4.2: Parking eType

Name	Description	DataType
Geographical name	a name by which a geographical location is known.	[] NLSTRING
Required Permit	what permit is needed to park	Concept
Total Spot Number	The total number of spots pertaining to this parking spot.	INTEGER
Available Spot Number	The number of spots available in this parking spot.	INTEGER
Maximum Allowed Height	Maximum allowed height for vehicles.	FLOAT
Maximum Parking Duration	Maximum allowed stay encoded as a ISO8601 duration	NLSTRING

explicit information using ontology. To make ATOM more compatible with an international standard, we use the components defined by INSPIRE directive on the transportation network. Our ontology clearly differentiates between transportation mean and mode. Currently, the terms are used interchangeably. This creates confusion while integrating multiple datasets. EU directive clearly distinguishes *mean* as the actual vehicle used for the mobility whereas *mode* is used to define as a way of travel i.e by air, road, and water. This distinction between components has to be explicit. For example, geographical locations (such as different transportation points: bus stop, taxi stand, and sea port) are the points from where a person use different transportation services (like bus, train, cable car etc). Figure 4.3 depicts five main classes such as location, event, structure, vehicle and organization which are aligned to the DOLCE [Guarino, 1998] top-level ontology.

Figure 4.3 shows the implemented model in *Protégé*. Here, the left panel shows the class hierarchy. There are five main classes: event, location, ve-

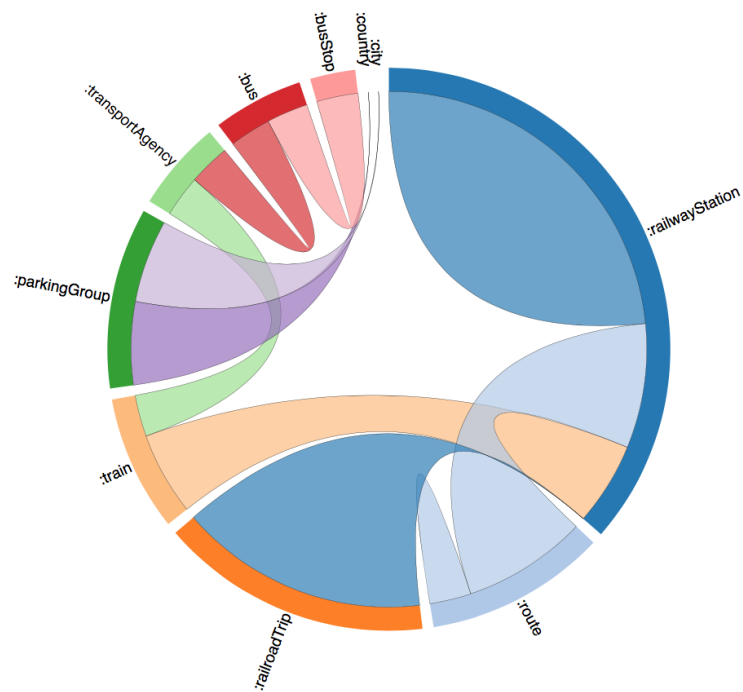


Figure 4.4: Class Relationship-Transportation

of *transportationArea* and is also a part of the *airport*. We preserve this type of semantic relations in our model which facilitates easy inference of the model. We categorize all the man made a *structure* like a *busStation*, *railwayStation*, *airport* etc under the class *transportationBuilding*. Moreover, to make our model comprehensible with more datasets, we gave major attention to not to use conflicting terms such as 'subways', 'underground' and 'tube'. The *Protégé* implementation of the ontology is available in web *Protégé* library⁵.

4.8 Language Level

We only included around 50 domain-specific concepts which were not available in the existing knowledge base. For example, concept like season ticket, permit active hours, total parking spot number etc.

4.9 Evaluation

To evaluate our model, we mapped two different datasets related to transportation. These data are collected from Open Street maps (OSM)⁶ available in Trentino, Italy and Open data Trentino portal⁷. The main criteria were based on the ability of ATOM to cover classes and attributes from a large number of dataset. The quality of our proposed model is evaluated based on four measurements: (a) completeness, (b) minimality, (c) understandability and (d) expressiveness, as proposed by [Akoka et al., 2007].

The result of the model evaluation is shown in Figure 4.5. The model was mapped to show that it is able to capture all the domain specific con-

⁵<https://goo.gl/fMgkeo>

⁶<https://www.openstreetmap.org/>

⁷<http://dati.trentino.it/dataset?groups=mobilita>

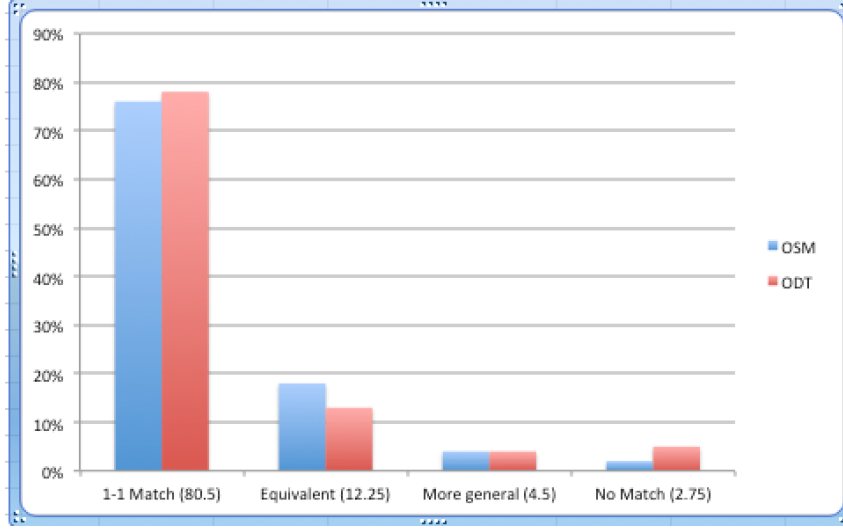


Figure 4.5: Statistics Transport

cepts (*Completeness*). For *completeness*, there was 80.5% *1-1 match*. *1-1 match* means when the target concept is found in defined vocabulary (for example railway-railway). Similarly, there was 12.25% *equivalent match* where *equivalent match* refers to a situation when there is a similar concept in the vocabulary (for example airdrome-airport). There was 4.5% *more general match*. The *more general match* refers to the immediate super-ordinate concept available than the concept available in the input schema (for example air travel-flight). Lastly, there was 2.75% *no match* between the concepts. We used less number of concepts to define our model (*Minimality*) as it mapped more than 1112 features that existed in OpenStreetMap (OSM) and other datasets to 120 concepts. The model also provided natural language description for all the concepts in order to make it expressive (*Expressiveness*) and understandable (*Understandability*) for common people.

In the second part we evaluated our model, based on the competency question mentioned in the Section 4.3. To answer the first competency question, we attach property *wheelchairAccess* with the class *transporta-*

tionPoint and set the datatype as Boolean. SPARQL query structure and the corresponding answer is shown in Figure 4.6.

```
PREFIX rdf:
<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX gtfs: <http://www.semanticweb.org/subhashis/
ontologies/2016/4/untitled-ontology-156#>
SELECT ?StopCode ?StopName ?wheelchairAccess
WHERE {
    ?busStop gtfs:stopCode ?StopCode.
    ?busStop gtfs:name ?StopName.
    ?busStop gtfs:wheelchairAccess ?wheelchairAccess.
}ORDERBY ?name LIMIT 5
```

StopName	wheelchairAccess	StopCode
"S.Ilario Sav"	^^<hl"true"^^<http://"	31171x"
"Lamar Ftm"	^^<ht"false"^^<http://"	22205z"
"Cognola Grezoni"	^."true"^^<http://"	24030z"
"Cognola Toresela"	^"true"^^<http://"	24015x"
"Cognola S.Vito"	^^."true"^^<http://"	24035x"

Figure 4.6: SPARQL query for wheelchair access

Similarly, to answer the second competency question (*Whether bicycle is allowed in the trip?*), we created *bikesAllowed* as a datatype property and attach it with the trip. Figure 4.7 shows the result from the query.

```
PREFIX rdf: <http://www.w3.org/1999/02
/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX gtfs:<http://www.semanticweb.org/subhashis
/ontologies/2016/4/untitled-ontology-156#>
SELECT ?TripName ?TripId ?bikesAllowed ?bar
WHERE { ?TripName gtfs:tripId ?TripId.
        ?TripName gtfs:bikesAllowed ?bikesAllowed.
        ?TripName gtfs:bar ?bar}
```

TripName	TripId	bikesAllowed	bar
Trento-Bassano_Del_Grappa	"5401"	^^"true"	^^<ht"false"
Trento-Bassano_Del_Grappa	"5405"	^^"true"	^^<ht"false"
Trento-Roma_Termini	"9477"	^^"false"	^^<ht"true"

Figure 4.7: SPARQL query for bikes allowed

Finally, to answer the third competency question (*What are the nearest facilities surrounding of a transportation stop?*), we used Apache Jena Fuseki ⁸ plugin. to allow running this type of spatial query. The answer of the query is shown in Figure 4.8.

```

PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX gtfs: <http://www.semanticweb.org/subhashis/
ontologies/2016/4/untitled-ontology-156#>
PREFIX spatial: <http://jena.apache.org/spatial#>
SELECT * {
    ?busStop spatial:nearby (46.08 11.06 1 'miles ');
    gtfs:name ?name
}

```

StopId	busStop	hotel
"21" ^^xsd:string	"Cognola Centro Civico"	"Hotel e Ristorante Villa Madruzzo"
"21" ^^xsd:string	"Cognola Centro Civico"	"Albermonaco"
"21" ^^xsd:string	"Cognola Centro Civico"	"Hotel Everest-Trento"

Figure 4.8: SPARQL query for near by services

The fourth and the fifth questions were similar to the above answered questions. So, we use other properties like *buyingProcess* along with the value such as *inTheBus* or *ticketOffice* and *petAllowed* with Boolean as a data type.

Figure 4.9 displaying knowledge graph (partial) for transportation model.

⁸<https://jena.apache.org/documentation/fuseki2/>

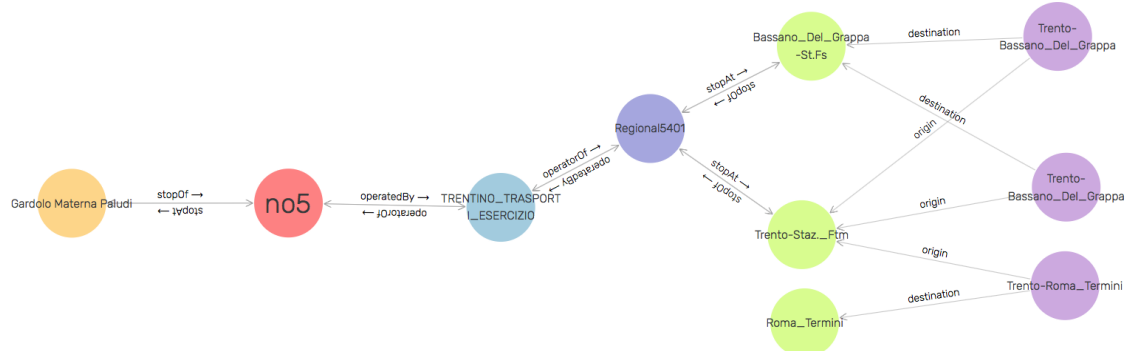


Figure 4.9: Transportation Knowledge Graph

GraphDB SPARQL Construct has been used to create the graph. It is to show that how different elements of the transportation domain related in the instance level. From the graph we can get information that *route number 5* has stop at *Gardolo Materna Paludi* and this bus service operated by *Trentino Transporti*, a transportation company. Trentino transporti also operated regional train service (e.g. *Regional5401*). Railway station *Trento Station FTM* and *Bassano Del Grappa St. FS* both are the stop of the regional train. Other information we can collect are the name of origin and destination of a train trip (e.g. *TrentoRoma Termini*).

Chapter 5

Healthcare Domain

“Health is the greatest gift,
contentment the greatest
wealth, faithfulness the best
relationship.” —*Buddha*

5.1 Domain Description

Semantic Interoperability in the health-care domain is of a great concern, where the main objective is to exchange health-related information with explicit meaning that is shared among different stakeholders or policymakers [Benson]. The main hurdle in achieving semantic interoperability is mainly because of difference in terminology used by various coding systems and also a lack of correlation among local coding systems and international coding systems. Ontology-based approach along with the standard controlled vocabulary helps tackle issues where the explicit formal specification is provided to interact with different systems [Héja et al., 2008]. Making health records meaningful will only be possible if we link the Electronic Health Record (EHR) to an authoritative clinical knowledge and then use natural language in the user interface as suggested by the IMIA working conference on clinical terminology which was held in 1984 [Benson]. This enables effec-

tive meaning-based retrieval. Figure 5.1 illustrated how healthcare domain intersects with other domains using the set diagram.

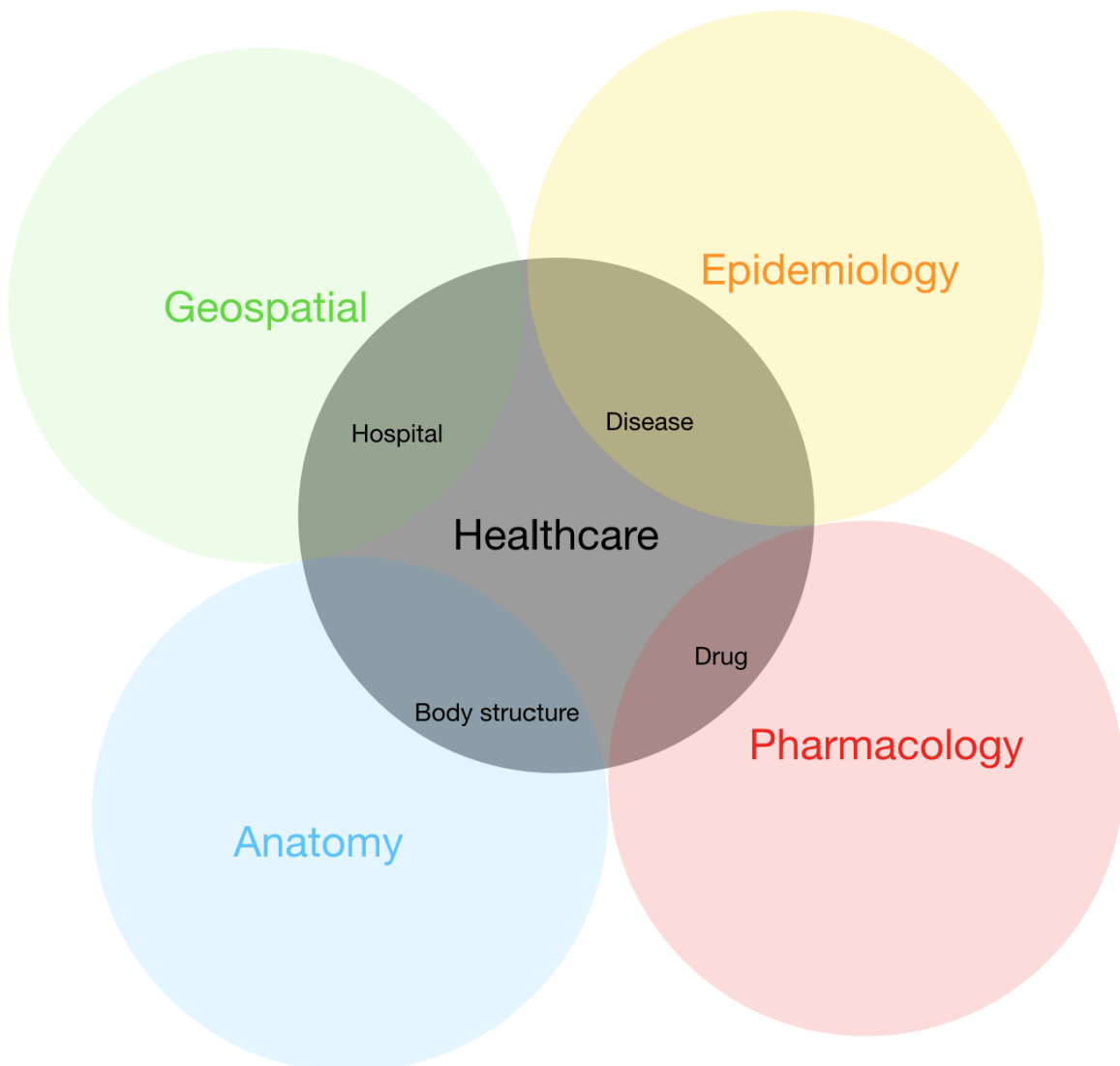


Figure 5.1: Healthcare Domain

5.2 Reference Scenarios

An organism working with healthcare and related data, such as a public healthcare data controller (e.g., NSS/eDRIS), needs to optimise their data preparation pipelines, in terms of time, effort, and competences invested into preparing data for research experiments. A key issue to address is data heterogeneity and the tedium associated with it in the daily work of data analysts (e.g., repeated solving of similar heterogeneity issues in order to maintain an acceptable quality of service to clients). Related issues are interoperability with data from external sources and interfacing with clients not familiar with local conventions and practices.

5.3 Generalized Queries

- Q1 all the patient (X) who diagnosis with condition (D) from country (Z)
- Q2 all the hospital (X) and healthcare (Y) clinic which has (Z) Clinical specialty.
- Q3 all the available drug product (X) for the disease (Y) with their daily dose (Z).
- Q4 all the patient (X) who visited Hospital (Y) from time t1 to time t2.

5.3.1 Identify the Domain Concepts

Electronic health record(EHR): EHR refers to an individual persons medical record in digital format. It may be made up of electronic medical records from many locations and/or sources. The EHR is a longitudinal electronic record of person health information generated by one or more encounters in any care delivery setting. Included in this information are

person demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data and radiology reports [Reich et al., 2017].

Electronic Medical Record (EMR): An electronic medical record is a computerized medical record created in an organization that delivers care, such as a hospital or outpatient setting. Electronic medical records tend to be a part of a local stand-alone health information system that allows storage, retrieval and manipulation of records. This document will reference EHR moving forward even if specific data source might internally use EMR definition [Reich et al., 2017].

Person: The Person class contains records that uniquely identify each patient in the source data who is time at-risk to have clinical observations recorded within the source systems [Reich et al., 2017].

Prescription: written instructions from a physician or dentist to a druggist concerning the form and dosage of a drug to be issued to a given patient [Miller and Fellbaum, 1998].

Visit: The visit class contains the spans of time a Person continuously receives medical services from one or more providers at a Care Site in a given setting within the health care system. Visits are classified into four categories: outpatient care, inpatient confinement, emergency room, and long-term care. Persons may transition between these settings over the course of an episode of care (for example, treatment of a disease onset) [Reich et al., 2017].

Death: The death class contains the clinical event for how and when a Person dies [Reich et al., 2017].

Specimen: The specimen class contains the records identifying biological samples from a person.

5.3.2 Query Collection

In principle, there are various ways to collect queries. For example, this can be done by analyzing the millions of user queries stored in the query logs of existing search portals or from interviewing the peoples. As our intended clients are from NHS (National Health Service) Scotland and Pharmaceutical company that is way we discussed with them to come up with all relevant queries they needed to get from the system.

5.4 State of the Art

5.4.1 Standards

Standards have evolved over many years to encompass more aspects of medicine, to cover them in more details, and to adapt as technology changes. Here, we describe more famous and widely accepted healthcare data standards.

International Statistical Classification of Diseases and Related Health Problems (ICD) : It is a medical classification list from the World Health Organization (WHO) [Organization, 2017]. It enlists codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases. It is available in 43 different languages used in 117 countries. ICD is updated every 10 years. ICD-9 is the 9th revision of the ICD (used in Italian data). ICD-10 is the current version and is the 10th revision of the ICD (used in Scottish data). ICD-10 is a major expansion capable of representing many more medical details.

The International Classification of Diseases for Oncology (ICD-O): It has been used for more than 35 years, principally in tumour or cancer

registries, for coding the site (topography)¹ and the histology (morphology)² of the neoplasm, usually obtained from a pathology report [Organization, 2013]. By agreement with the College of American Pathologists, the morphology section of ICD-O is incorporated into the Systematized Nomenclature of Medicine (SNOMED) classification as the neoplasm section of the morphology field. New version is ICD-O 3. This updated version of ICD-O-3 (ICD-O-3 First Revision, or ICD-O-3.1) includes the new terms, codes, synonyms, related terms, morphology, and behaviour code changes from the WHO Blue Books published between 2007 and 2010 on tumours of haematopoietic and lymphoid tissues⁸, the central nervous system⁹, and the digestive system¹⁰ [Fritz et al., 2000].

SNOMED-CT: It is a comprehensive reference terminology that supports both general and more specific concepts [International, 2017]. Concepts are a basic component of a SNOMED-CT and have a clinical meaning. They are identified by a unique nine digit numeric concept ID (e.g. 17373004) and a unique human-readable (e.g. Bilateral kidneys) fully specified name. Each concept is defined by a set of attribute-value pairs (relationship) which make it distinct from all other concepts.

Current Procedural Terminology (CPT): It is maintained and updated annually by the American Medical Association to classify all medical procedures [Association, 2007]. And it is required for virtually all the billing and reimbursement process. CPT code is divided into three categories. Category I codes are widely performed procedures and are 5 digits long divided into sections for anesthesiology, surgery, radiology, pathology, laboratory medicine, and medicine. Category II codes are for the collection of quality and performance metrics and are of 4 digits. Category III codes are for new or experimental procedures and also have 4 digits. For each

¹<http://codes.iarc.fr/topography>

²<http://codes.iarc.fr/codegroup/2>

code, there are full, medium, and short names or descriptions. For example, flu vaccine, influenza vaccine. These are used for different purposes and provide varied levels of details as needed. Given it's used for billing, a CPT code may need to provide details necessary to determine the proper charge.

OPCS-4: The OPCS-4³ Classification of Interventions and Procedures (OPCS-4) is a statistical classification for clinical coding of hospital interventions and procedures undertaken by the National Health Service (NHS) [of General Practitioners et al., 1991]. It is an approved NHS Fundamental Information Standard. The classification is mandatory for use by health-care providers to support various forms of data collections for operational and secondary uses. It is mandatory for Admitted Patient Care Commissioning Data Sets (CDS). The requirements for data sets and related definitions are specified in the NHS Data Model and Dictionary. The OPCS-4 classification is updated to accurately reflect current NHS clinical practice through the addition of new content as necessary. Key classification characteristics are retained to maintain consistency and comparability across OPCS-4 versions and statistical data.

Logical Observation Identifiers Names and Codes (LOINC): It is a common language (a set of identifiers, names, and codes) for identifying health measurements, observations, and documents. LOINC is a rich catalog of measurements, including laboratory tests, clinical measures like vital signs and anthropomorphic measures, standardized survey instruments, and more [Reichert et al., 2002]. LOINC also contains codes for collections of these items, such as panels, forms, and documents. For each *concept*, LOINC contains many other rich details, such as synonyms, units of measure, and carefully crafted descriptions.

³<http://www.datadictionary.nhs.uk/>

National Drug Code (NDC)⁴: It is a USA specific standard for medications maintained by the Food and Drug Administration (FDA) [Food and , FDA]. It consists of 10 digits, three segment structure to indicate the Labeler or Vendor, the Drug, and the Packaging.

The British National Formulary (BNF) is a United Kingdom (UK) pharmaceutical reference book that contains a wide spectrum of information and advice on prescribing and pharmacology, along with specific facts and details about many medicines available on the UK National Health Service (NHS) [Association et al., 1999]. Information within the BNF includes indication(s), contraindications, side effects, doses, legal classification, names and prices of available proprietary and generic formulations, and any other notable points.

WHO Adverse Reactions Terminology (WHO-ART): It is a four-level hierarchical terminology, which begins at the body system/organ level classes. These classes consist of broad grouping terms, which consist of more specific preferred terms [Centre, 2010]. WHO also contains commonly used terms, called included terms, that act as entry terms for the preferred terms.

MedDRA or Medical Dictionary for Regulatory Activities: It is a clinically validated international medical terminology dictionary (and thesaurus) used by regulatory authorities in the pharmaceutical industry during the regulatory process, from pre-marketing to post-marketing activities, and for data entry, retrieval, evaluation, and presentation [Brown et al., 1999]. MedDRA originally available in English and Japanese, MedDRA is now also translated into Chinese, Czech, Dutch, French, German, Hungarian, Italian, Portuguese and Spanish.

RxNorm: It provides normalized names for clinical drugs and links its names to many of the drug vocabularies commonly used in pharmacy man-

⁴<https://www.fda.gov/Drugs/InformationOnDrugs/ucm142438.htm>

agement and drug interaction software, including those of First Databank, Micromedex, MediSpan, Gold Standard Drug Database, and Multum. By providing links between these vocabularies, RxNorm can mediate messages between systems not using the same software and vocabulary [Liu et al., 2005].

Fast Healthcare Interoperability Resources (FHIR, pronounced “fire”):

It is a draft standard describing data formats and elements (known as “resources”) and an Application Programming Interface (API) for exchanging Electronic health records. The standard was created by the Health Level Seven International (HL7) healthcare standards organization. FHIR is built on the previous data format standards from HL7⁵, like HL7 version 2.x and HL7 version 3.x. It is easier to implement because it uses a modern web-based suite of API technology, including an HTTP-based RESTful protocol, HTML and Cascading Style Sheets for user interface integration, a choice of JSON or XML for data representation, and Atom for results.

ASTM CCR : The Continuity of Care Record (CCR) is a core data set of the most relevant administrative, demographic, and clinical information facts about a patients healthcare, covering one or more healthcare encounters [Standard, 2005]. It provides a means for one healthcare practitioner, system, or setting to aggregate all of the pertinent data about a patient and forward it to another practitioner, system, or setting to support the continuity of care. The CCR data set includes a summary of the patients health status (for example, problems, medications, allergies) and basic information about insurance, advance directives, care documentation, and the patients care plan. It also includes identifying information and the purpose of the CCR.

Reference Information Model (RIM): It is the cornerstone of the

⁵HL7 is a non-profit organization involved in the development of international healthcare informatics interoperability standards known as Health Level 7

HL7 Version 3 development process [Eggebraaten et al., 2007]. An object model created as part of the Version 3 methodology, the RIM is a large, pictorial representation of the HL7 clinical data (domains) and identifies the life cycle that a message or groups of related messages will carry. It is a shared model between all domains and, as such, is the model from which all domains create their messages. The RIM is an ANSI approved standard.

Observational Medical Outcomes Partnership (OMOP): The Observational Medical Outcomes Partnership (OMOP)⁶ was a public-private partnership established to inform the appropriate use of observational healthcare databases for studying the effects of medical products [Reich et al., 2017]. Over the course of the 5-year project and through its community of researchers from industry, government, and academia, OMOP successfully achieved its aims to: 1) conduct methodological research to empirically evaluate the performance of various analytical methods on their ability to identify true associations and avoid false findings, 2) develop tools and capabilities for transforming, characterizing, and analyzing disparate data sources across the health care delivery spectrum, and 3) establish a shared resource so that the broader research community can collaboratively advance the science.

5.4.2 Applications

The UMLS Semantic Network⁷: The Semantic Network consists of (1) a set of broad subject categories, or Semantic Types, that provide a consistent categorization of all concepts represented in the UMLS Metathesaurus, and (2) a set of useful and important relationships, or Semantic Relations, that exist between Semantic Types. This section of the docu-

⁶<http://omop.org/CDM>

⁷<https://semanticnetwork.nlm.nih.gov>

mentation provides an overview of the Semantic Network, and describes the files of the Semantic Network. Sample records illustrate structure and content of these files.

The SPECIALIST NLP Tools⁸: The SPECIALIST Natural Language Processing (NLP) Tools have been developed by the The Lexical Systems Group of The Lister Hill National Center for Biomedical Communications to investigate the contributions that natural language processing techniques can make to the task of mediating between the language of users and the language of online biomedical information resources. The SPECIALIST NLP Tools facilitate natural language processing by helping application developers with lexical variation and text analysis tasks in the biomedical domain. The NLP Tools are open source resources distributed subject to these terms and conditions.

The RxClass⁹ Browser (See Figure 5.2) is a web application for exploring and navigating through the class hierarchies to find the RxNorm drug members associated with each class. RxClass links drug classes of several drug sources including ATC, MeSH, NDF-RT and FDA/SPL to their RxNorm drug members (ingredients, precise ingredients and multiple ingredients). RxClass allows users to search by class name or identifier to find the RxNorm drug members or, conversely, search by RxNorm drug name or identifier to find the classes that the RxNorm drug is a member of.

Aggregated Patient Data: 1upHealth's¹⁰ provider application helps aggregate patient data from external health systems into one place. Providers can view the data sources that patients have connected via the 1upHealth patient application (<https://1uphealth.care/patient>). Data is presented in an easy to understand patient timeline which supports demographics,

⁸<https://lexsrv3.nlm.nih.gov/Specialist/Home/index.html>

⁹<https://mor.nlm.nih.gov/RxClass/>

¹⁰<https://1uphealth.care>

class: Lung Neoplasms / id: N0000001883 / class type: DISEASE / [show context](#) Print

20 RxNorm generic drugs for **may_treat** in NDFRT / [similar classes](#)

Type	RXCUI	RxNorm Name	Relation	All classes
IN	40048	Carboplatin	DIRECT	Show
IN	2105	Carmustine	DIRECT	Show
IN	2418	Cholecalciferol	DIRECT	Show
IN	2555	Cisplatin	DIRECT	Show
IN	3002	Cyclophosphamide	DIRECT	Show
IN	72962	docetaxel	INDIRECT	Show
IN	3639	Doxorubicin	DIRECT	Show
IN	4179	Etoposide	DIRECT	Show
IN	328134	gefitinib	DIRECT	Show

Figure 5.2: RxClass Browser

medications, labs, conditions, and history.

M*Modal Fluency Direct¹¹: It is a next-generation, all-in-one speech recognition solution that enables physicians of any medical specialty to conversationally create, review, edit and sign clinical notes directly within EHR templates. It leverages M*Modal Speech Understanding™, which is far more than just voice-to-text technology as it includes Natural Language Understanding (NLU) technology for contextual understanding of the physician narrative to help improve accuracy.

Human API Human API¹² is a platform that allows users to securely share their health data with developers of health applications and systems. Our data network includes activity data recorded by pedometers, blood pressure measurements captured by digital cuffs, medical records from hospitals, and more. Human API takes care of the data synchronizations with third party data sources, user management for your user identities across all data sources, and processes and normalizes the data

¹¹<https://mmodal.com/wp-content/uploads/2016/02/MModal-Fluency-Direct-Brochure.pdf>

¹²<http://hub.humanapi.co/v1.1/docs/architecture>

over a secure, HIPAA compliant, RESTful API. Architecture of Human API shown below Figure 5.3.

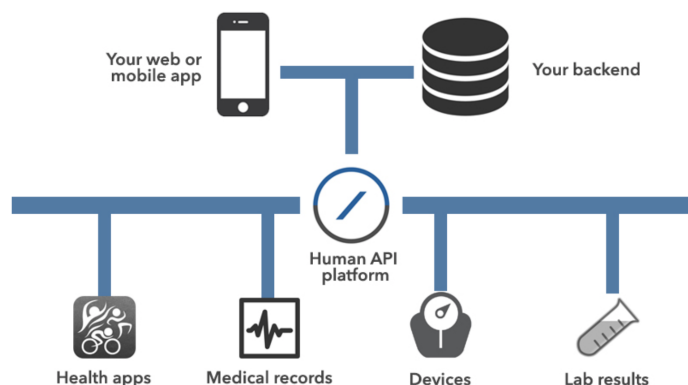


Figure 5.3: Human API

DrugBank¹³: The DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information (see Figure 5.4). The database contains 8261 drug entries including 2021 FDA-approved small molecule drugs, 233 FDA-approved biotech (protein/peptide) drugs, 94 nutraceuticals and over 6000 experimental drugs[Law et al., 2014]. These DrugBank datasets are released under a Creative Commons Attribution-NonCommercial 4.0 International License. They can be used freely in your non-commercial application or project.

SMART is An App Platform for Healthcare. SMART¹⁴ Health IT is an open, standards based technology platform that enables innovators to create apps that seamlessly and securely run across the healthcare system. Using an electronic health record (EHR) system or data warehouse that supports the SMART standard, patients, doctors, and healthcare practi-

¹³<https://www.drugbank.ca/about>

¹⁴<https://smarthealthit.org/an-app-platform-for-healthcare/about/>



Identification	
Name	Warfarin
Accession Number	DB00682 (APRD00341)
Type	Small Molecule
Groups	Approved
Description	Warfarin is an anticoagulant drug normally used to prevent blood clot formation as well as migration. Although originally marketed as a pesticide (d-Con, Rodex, among others), Warfarin has since become the most frequently prescribed oral anticoagulant in North America. Warfarin has several properties that should be noted when used medically, including its ability to cross the placental barrier during pregnancy which can result in fetal bleeding, spontaneous abortion, preterm birth, stillbirth, and neonatal death. Additional adverse effects such as necrosis, purple toe syndrome, osteoporosis, valve and artery calcification, and drug interactions have also been documented with warfarin use. Warfarin does not actually affect blood viscosity, rather, it inhibits vitamin-k dependent synthesis of biologically active forms of various clotting factors in addition to several regulatory factors.
Structure	 <div>MOL SDF 3D-SDF PDB SMILES InChI View 3D Structure</div>
Synonyms	4-Hydroxy-3-(3-oxo-1-phenylbutyl)coumarin Coumafene  Zocoumarin

Figure 5.4: DrugBank Browser

tioners can draw on this library of apps to improve clinical care, research, and public health.

5.5 Informal Modeling Phase

5.5.1 Pilot Reference Datasets

We analyzed four datasets, as part of our preliminary study to understand how information is stored in those datasets. Geographical dataset (see Table 5.1) contain information about data zone (it is a small geographical area) and their ranking in terms of a social index (i.e. SIMD). Dataset also provides information regarding material disadvantage in the population in terms of lack of car ownership. The Carstairs index¹⁵ has been produced based on census data. Scottish Morbidity Records (SMR) and National Records of Scotland (NRS) dataset contains sensitive personal information of a patient such as ethnic group, marital status (see Table 5.2.

Prescription dataset contains (i.e. PIS Variable) all sensitive information (i.e. personal information) related to a patient and prescriber. Demographic information available in the dataset are gender, address, date

¹⁵<http://www.isdscotland.org/products-and-Services/GPD-Support/deprivation/carstairs/index.asp?Co=Y>

Table 5.1: Healthcare Dataset 1

DatasetName: Geographical Variables				
Attribute:	character postcode	SIMD 2016 rank	Carstairs score	SIMD score
Class:	Data zone, Council Area, Intermediate Zone			

Table 5.2: Healthcare Dataset2

DatasetName: SMR & NRS	
Attribute	ETHNIC_GROUP, MARITAL_STATUS, DISCHARGE_DATE, LENGTH_OF_STAY
Class:	Patient

of birth and occupation. Analysis also helps to understand coding system used to codifying disease names. They are ICD-10, ICD-9, ICD-O-2.

5.5.2 ExER Model

Extended Entity Relationship (ExER) model for Healthcare shown in Figure 5.5. The patient entity has community health index (CHI) number as an identifying attribute and age (at time of visit). For the privacy reason, we separated all personal information of a patient from a person and connected with the *roleOf* relation. So that user can able to access personal information of a patient if they have right permission. Person entity has all demographic information such as date of birth, ethnic group, and postal address. With the relation *countryofBirth*, we can collocate all person who born in the same country. The country entity has the attribute such as ISO code, population, currency name etc. Visit entity store all information of the patient visit to a health encounter site. Hospital entity has an attribute address (a structure attribute) which store information about house number, street name, street name, postcode, country, and city.

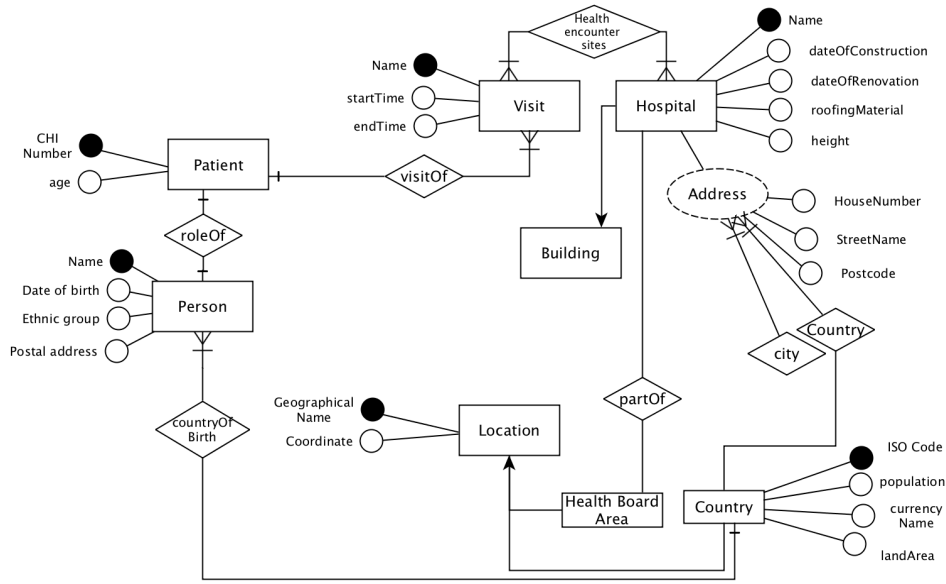


Figure 5.5: ExER Model for Healthcare

5.6 Health eTypes

We developed health eTypes model based on the design principles instructed by the OMOP common data model (CDM) [Reich et al., 2017]. The OMOP CDM is designed to include all observational health data elements (experiences of the patient receiving health care) that are relevant for analysis use cases to support the generation of reliable scientific evidence about disease natural history, healthcare delivery, effect of medical interventions, the identification of demographic information, health care interventions and outcomes. Therefore, the CDM is designed to store observational data to allow for research, under the following principles:

- Suitability for purpose: The CDM aims at providing data organized in a way optimal for analysis, rather than for the purpose of operational needs of health care providers or payers.
- Data protection: All data that might jeopardize the identity and protection of patients, such as names, precise birthdays etc. are limited.

Exceptions are possible where the research expressly requires more detailed information, such as precise birth dates for the study of infants.

- Design of domains: The domains are modeled in a entity-centric relational data model, where for each record the identity of the person and a date is captured as a minimum.
- Rationale for domains: Domains are identified and separately defined in an Entity-relationship model if they have an analysis use case and the domain has specific attributes that are not otherwise applicable. All other data can be preserved as an observation in an entity-attribute-value structure.
- Standardized Vocabularies: To standardize the content of those records, the CDM relies on the Standardized Vocabularies containing all necessary and appropriate corresponding standard healthcare concepts.
- Reuse of existing vocabularies: If possible, these concepts are leveraged from national or industry standardization or vocabulary definition organizations or initiatives, such as the National Library of Medicine, the Department of Veterans Affairs, the Center of Disease Control and Prevention, National Health Service etc.
- Maintaining source codes: Even though all codes are mapped to the Standardized Vocabularies, the model also stores the original source code to ensure no information is lost.
- Technology neutrality: The CDM does not require a specific technology. It can be realized in any relational database, such as Oracle, SQL Server etc., or as SAS analytical datasets.
- Scalability: The CDM is optimized for data processing and computational analysis to accommodate data sources that vary in size, includ-

ing databases with up to hundreds of millions of persons and billions of clinical observations.

- Backwards compatibility: All changes from previous CDMs are clearly delineated. Older versions of the CDM can be easily created from this CDMv5, and no information is lost that was present previously.

Prescription eType (see Table 5.3) collected all vital information of a patient. It store the time when a prescription is issued and also when drugs is dispense from the pharmacy.

Table 5.3: Prescription eType

Name	Description	DataType
Name	The name by which an entity is known	[] NLSTRING
Identifier	a symbol that establishes the identity of the one bearing it	STRING
Patient	a person who requires medical care	Patient
Prescription drug	a drug that is available only with written instructions from a doctor or dentist to a pharmacist	Drug
Prescription date	the date on which a prescription is issued.	DATE
Dispense date	the date on which drugs dispense	DATE

Patient eType (see Table 5.4) contain only Patient Identification number unlike the OMOP CDM [Reich et al., 2017] model where all personal information as well as patient is store under person class. But because of privacy reason we decided to separate sensitive personal information from the patient eType and keep it in the person eType.

Visit eType (see Table 5.5) is contain information about all visits of patient to a healthcare centre. An inpatient admission marks the start of an inpatient episode. The patient undergoes the full admission procedure and is accepted by the hospital, the specialty and the consultant for inpatient

Table 5.4: Patient eType

Name	Description	DataType
Identifier	a symbol that establishes the identity of the one bearing it	STRING
Finding related to biological sex	the properties that distinguish organisms on the basis of their reproductive roles	Concept
CHI Number	Community Health Index Number	SSTRING
Visit	person visit to the healthcare centre	Visit
Prescription	written instructions from a physician	Prescription

care. The full admission procedure may be defined as the completion of all registration documents including the recording of the patient's name in the admission register or system. The patient may be admitted from: a hospital with in the same health board, a hospital situated in different health board or outpatients.

Table 5.5: Visit eType

Name	Description	DataType
Identifier	a symbol that establishes the identity of the one bearing it	STRING
Patient	a person who requires medical care	Patient
Health encounter sites	Healthcare centre where patient receive care	Hospital

Person eType (see Table 5.6) is collect all personal information of a patient who seek healthcare support or who at least visited a healthcare centre for medical checkup. It is our design decision to keep personal information of a patient separate from patient profile because of the privacy policy of the health board. As we can see there is a CHI number through which we can access personal informal of a patient if required and have authority and permission.

Table 5.6: Person eType

Name	Description	Data Type
Identifier	a symbol that establishes the identity of the one bearing it	STRING
Person name	a name by which a person is known	STRING
Finding related to biological sex	the properties that distinguish organisms on the basis of their reproductive roles	Concept
Employment status	current employment status of the person	Concept
Ethnic group	an ethnic quality or affiliation resulting from racial or cultural ties	Concept
CHI Number	Community Health Index Number	STRING
Date of birth	the date on which the person born	DATE
Birth place	the place where someone was born	Location
Country of residence	the country where the person has residency	Country
Address	written directions for finding some location	Address

5.7 Schema Level

5.7.1 Defined Classes

Figure 5.6 showing the class hierarchy for the healthcare domain. On the right-hand side of the figure, it displaying all subclasses of the event class. It includes medical procedure which records all information related to medical treatment or laboratory test of a patient. patientVisit store all visiting information of a patient, episode is used for collocate all independent but interconnected events related to a patient visit who changes different wards of the same hospital without discharge from the hospital.

The intermediate class hierarchy is occupied by the domain concept. All domain concepts are generated to capture the healthcare domain. Figure 5.6 depicted the domain hierarchy of the model. The classes with green square bracket are taken from OMOP standard [Reich et al., 2017]. Other classes are created to accommodate domain concept from SNOMED-CT.

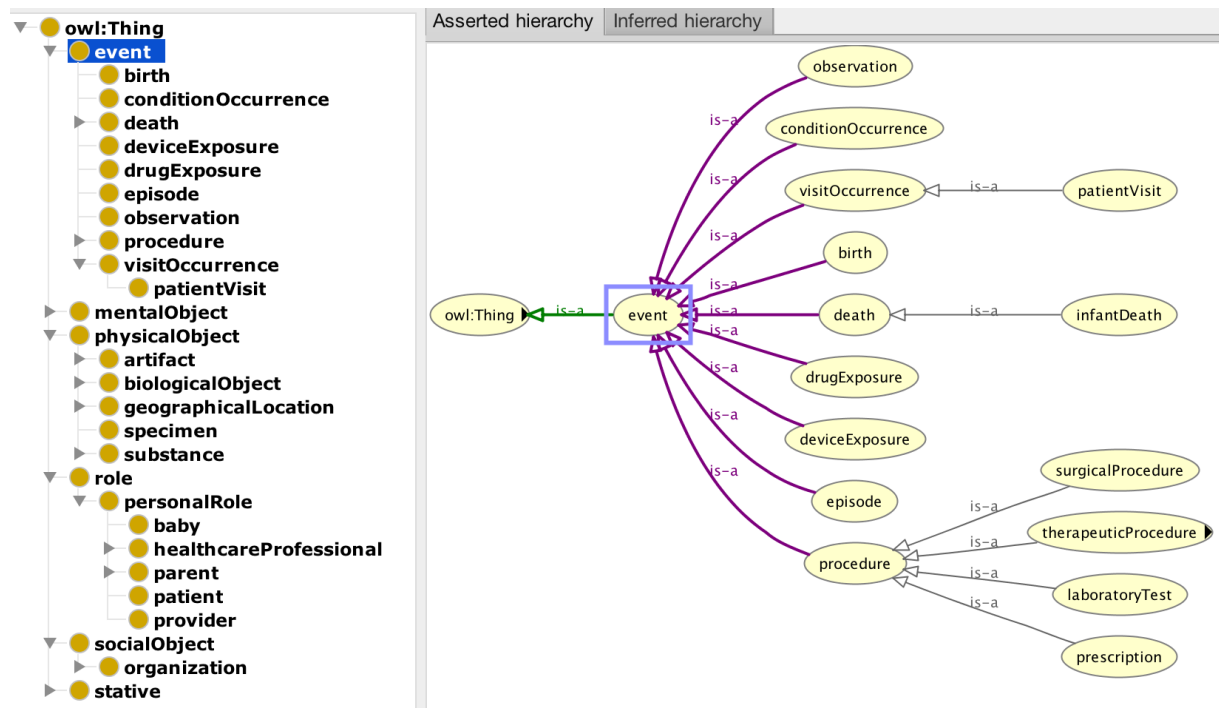


Figure 5.6: Class Hierarchy of Healthcare

Here we provided little description of all domain concept taken from OMOP data model [Reich et al., 2017]:

Death: The death class contains the clinical event for how and when a Person dies. A person can have up to one record if the source system contains evidence about the Death, such as: 1) Condition Code in the Header or Detail information of claims 2) Status of enrollment into a health plan 3) Explicit record in EHR data.

Visit occurrence: The visit occurrence class contains the duration of time a Person continuously receives medical services from one or more providers at a Care Site (i.e. Hospital) in a given setting within the health care system (e.g. NHS system). Visits are classified into four categories: outpatient care, inpatient, emergency, and long-term care. Persons may transition between these category over the course of an episode of care (for

example, treatment of a disease onset).

Device exposure: The device exposure class captures information about a persons exposure to a foreign physical object (or device) or instrument that which is used for diagnostic or therapeutic purposes through a mechanism beyond chemical action. Devices include implantable objects (e.g. pacemakers, artificial joints), medical equipment and supplies (e.g. bandages, crutches, syringes), other instruments used in medical procedures (e.g. sutures, defibrillators) and material used in clinical care (e.g. adhesives, body material, dental material, surgical material).

Drug exposure: The drug exposure class captures records about the utilization of a Drug when ingested or otherwise introduced into the human body. A Drug is a biochemical substance formulated in such a way that when administered to a Person it will exert a certain physiological effect. Drugs include prescription and over-the-counter medicines, vaccines, and large-molecule biologic therapies. Radiological devices ingested or applied locally do not count as Drugs. Drug Exposure is inferred from clinical events associated with orders, prescriptions written, pharmacy dispensings, procedural administrations, and other patient-reported information, for example: in the *prescription* section of an EHR, the *Medication* section of an EHR or Drugs administered as part of a Procedure (or therapy), such as chemotherapy or vaccines.

Procedure occurrence: The procedure occurrence class contains records of activities or processes ordered by, or carried out by, a health-care provider on the patient to have a diagnostic or therapeutic purpose. Procedures are present in various data sources in different forms with varying levels of standardization. For example: 1) Medical Claims include procedure codes that are submitted as part of a claim for health services rendered, including procedures performed. 2) EHR that capture procedures as orders.

Class relationship has been shown in Figure 5.7. It showing the depen-

dencies between 9 classes. In particular, prescription has 1545 relations in total which include 545 incoming relation and 1000 outgoing relations.

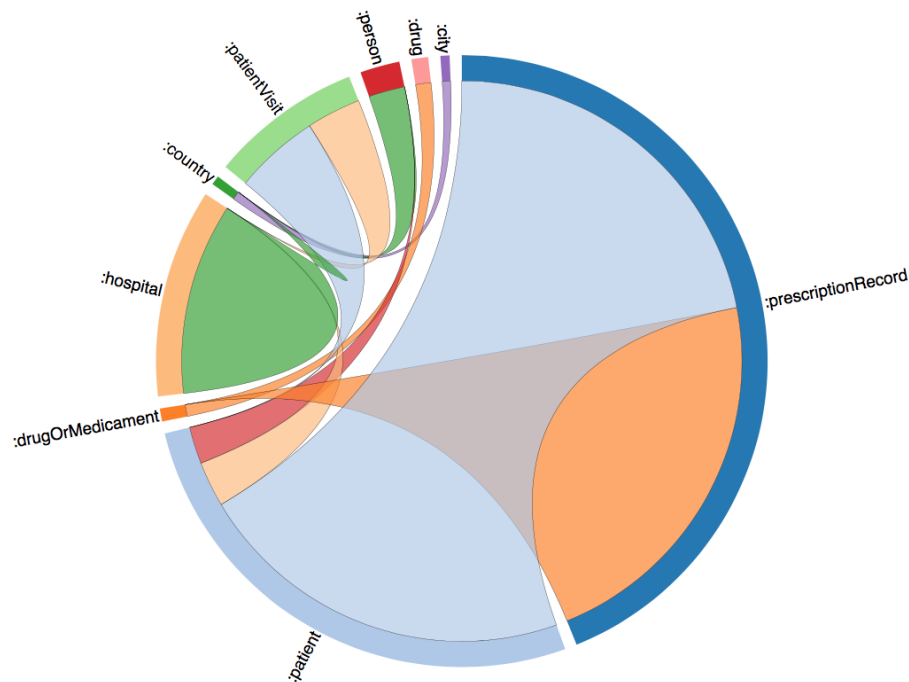


Figure 5.7: Class Relationship-Healthcare

5.7.2 Alignment with top level ontology

The top level of the class hierarchy is preoccupied with the upper-level concepts from foundational ontologies (e.g. DOLCE [Guarino, 1998]). For example, event, stative, physical object, mental object and social object all these concepts are defined in the DOLCE ontology. All other classes are descended from this root class. This means that the root class is a supertype of all other classes. The direct subtypes of the root class are referred to as ‘Top Level Classes’.

5.8 Language Level

We imported 338000 concepts (English and Spanish) from SNOMED-CT. We considered *fully specified name*¹⁶ without parenthesis as the entry term. For example, *Primary malignant neoplasm of lung* will be the entry term and not *Primary malignant neoplasm of lung (disorder)*. We also decided to use SNOMED-CT code as a word lemma. For example, code for the lung cancer i.e. *SCTID-93880001* will be treated as a word.

For the ranking of words, we chose first one as an entry term and other as synonymy¹⁷ word and rank them based on appearance on the dataset. And then add code as a word. In our integrated language resource, it will appear as

Example 1: Primary malignant neoplasm of lung, Lung cancer, SCTID:93880001

5.9 Evaluation

For healthcare domain, we only perform query evaluation. Compare with other datasets yet to be done. Here, we mentioned some of the SPARQL queries to show that model can able to answer the necessary question which might be useful for analytics purpose and further research.

Q1 Select all female patient who visited hospital and prescribe quantity.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX shib: <http://www.semanticweb.org/subhashis/
\\ontologies/2017/4/untitled-ontology-446#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT DISTINCT ?substanceName ?Age ?prescribedQuantity
WHERE {
    ?patientVisit shib:ageInYears ?Age.
    ?patientVisit shib:visitOfPatient ?patient.
```

¹⁶Each SNOMED-CT concept has at least one Fully Specified Name (FSN) intended to provide an unambiguous way to name a concept. The purpose of the FSN is to uniquely describe a concept and clarify its meaning.

¹⁷A synonym is a term with the same meaning as another term. Terms that are synonyms are said to be synonymous.

```

    ?patient shib:roleOf ?person.
    ?person shib:forename ?PersonForename.
    ?person shib:surname ?PersonSurname.
  OPTIONAL {?person shib:findingRelatedToBiologicalSex ?Sex.}
    ?person shib:countryOfBirth ?Country.
    ?Country rdfs:label ?CountryOfBirth.
    ?patient rdfs:label ?PatientUPI.
    ?patient shib:hasPrescription ?Prescription.
    ?Prescription rdfs:label ?PrescriptionID.
    ?Prescription shib:prescribedQuantity ?prescribedQuantity.
    ?Prescription shib:prescriptionDrug ?drugSubstance.
    ?drugSubstance rdf:type shib:drugOrMedicament.
    ?drugSubstance rdfs:label ?substanceName.
    ?drugSubstance shib:legalDrugProduct ?drugProduct.
    ?drugProduct rdfs:label ?ProductName.
  FILTER regex(?Sex, "Female")
}

```

Q2 Select all male patient who visited hospital and prescribe quantity.

```

PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX shib: <http://www.semanticweb.org/subhashis/
  \\ontologies/2017/4/untitled-ontology-446#>
PREFIX rdf: <http://www.w3.org/1999/02/
  \\22-rdf-syntax-ns#>
SELECT DISTINCT    ?substanceName ?Age ?prescribedQuantity
  WHERE {
    ?patientVisit shib:ageInYears ?Age.
    ?patientVisit shib:visitOfPatient ?patient.
    ?patient shib:roleOf ?person.
    ?person shib:forename ?PersonForename.
    ?person shib:surname ?PersonSurname.
  OPTIONAL {?person shib:findingRelatedToBiologicalSex ?Sex.}
    ?person shib:countryOfBirth ?Country.
    ?Country rdfs:label ?CountryOfBirth.
    ?patient rdfs:label ?PatientUPI.
    ?patient shib:hasPrescription ?Prescription.
    ?Prescription rdfs:label ?PrescriptionID.
    ?Prescription shib:prescribedQuantity ?prescribedQuantity.
    ?Prescription shib:prescriptionDrug ?drugSubstance.
    ?drugSubstance rdf:type shib:drugOrMedicament.
    ?drugSubstance rdfs:label ?substanceName.
    ?drugSubstance shib:legalDrugProduct ?drugProduct.
    ?drugProduct rdfs:label ?ProductName.
  }

```

```
FILTER regex (?Sex, "Male")
}
```

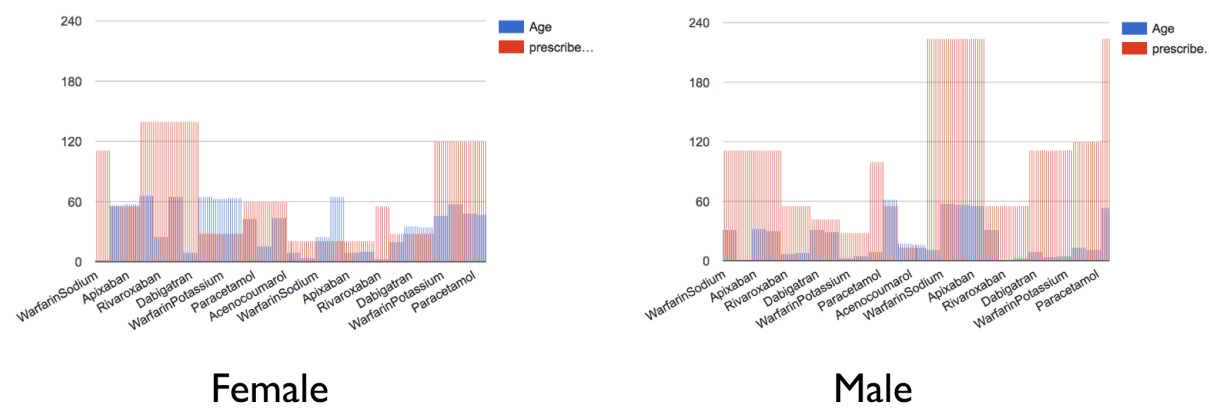


Figure 5.8: comparison chart showing prescribe quantity of Drug X in respect with their age

Q3 Retrieve the all prescribed drugs information of a patient along with the name of drugs manufacturer and patient personal information.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX shib: <http://www.semanticweb.org/subhashis/
ontologies/2017/4/untitled-ontology-446#>
PREFIX rdf: <http://www.w3.org/1999/
02/22-rdf-syntax-ns#>
PREFIX : <http://www.semanticweb.org/subhashi/
ontologies/2017/4/untitled-ontology-446#>
SELECT ?PatientUPI ?PersonForename ?PersonSurname ?CountryOfBirth
?PrescriptionID ?substanceName ?ProductName
WHERE { ?patient shib:roleOf ?person .
        ?person shib:forename ?PersonForename .
        ?person shib:surname ?PersonSurname .
        ?person shib:countryOfBirth ?Country .
        ?Country rdfs:label ?CountryOfBirth .
        ?patient rdfs:label ?PatientUPI .
        ?patient shib:hasPrescription ?Prescription .
        ?Prescription rdfs:label ?PrescriptionID .
        ?Prescription shib:prescriptionDrug ?drugSubstance .
        ?drugSubstance rdf:type shib:drugOrMedicament .
        ?drugSubstance rdfs:label ?substanceName .
        ?drugSubstance shib:legalDrugProduct ?drugProduct .
        ?drugProduct rdfs:label ?ProductName .
```

	PatientUPI ⇅	PersonForename ⇅	PersonSurname ⇅	CountryOfBirth ⇅	PrescriptionID ⇅	substanceName ⇅	ProductName ⇅
1	"1234567903"@en	Carl	Brown	"France"@en	"11238"@en	"WarfarinSodium"@en	"WarfarinSodium1mg"@en
2	"1234567903"@en	Carl	Brown	"France"@en	"11238"@en	"Apixaban"@en	"WarfarinSodium1mg"@en
3	"1234567903"@en	Carl	Brown	"France"@en	"11238"@en	"Rivaroxaban"@en	"WarfarinSodium1mg"@en

Figure 5.9: Patient personal information along with the prescribe drugs

Q4 Show all connection of a patient in a knowledge graph using SPARQL construct.

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX shib:
```

```

<http://www.semanticweb.org/subhashis/ontologies/
\\2017/4/untitled-ontology-446#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
CONSTRUCT { ?patientVisit shib:visitOfPatient ?patient.
              ?patient shib:roleOf ?person.
              ?person shib:countryOfBirth ?Country.
              ?patient shib:hasPrescription ?Prescription.
              ?Prescription shib:prescriptionDrug ?drugSubstance.
              ?drugSubstance shib:legalDrugProduct ?drugProduct.

              }
WHERE {
  ?patientVisit shib:visitOfPatient ?patient.
  ?patient shib:roleOf ?person.
  ?person shib:forename ?PersonForename.
  ?person shib:surname ?PersonSurname.
  ?person shib:countryOfBirth ?Country.
  ?Country rdfs:label ?CountryOfBirth.
  ?patient rdfs:label ?PatientUPI.
  ?patient shib:hasPrescription ?Prescription.
  ?Prescription rdfs:label ?PrescriptionID.
  ?Prescription shib:prescriptionDrug ?drugSubstance.
  ?drugSubstance rdf:type shib:drugOrMedicament.
  ?drugSubstance rdfs:label ?substanceName.
  ?drugSubstance shib:legalDrugProduct ?drugProduct.
  ?drugProduct rdfs:label ?ProductName.
}

```

The knowledge graph (see Figure 5.10) to show how the patient 1234567928 (PatientId) is related with the person 38 (personID) with role of relation. Total seven prescriptions have been prescribed to the patient. All prescription prescribed a same drugs which is Warfarin sodium. And for that generic drug, there is four product available in the market.



Figure 5.10: Partial view of Healthcare Knowledge Graph

5.10 Issues

Mapping from ICD-9 to ICD-10: Difference version of the same standard (e.g. ICD-9 (Old version) and ICD-10 (New version)) can create many challenges for mapping even for the expert because of the granularity. For example, ICD-10-CM has 68,000 codes, compared to 13,000 in ICD-9-CM, according to the Centers for Medicare & Medicaid Services (CMS)¹⁸. Figure 5.11 shows a comparison between ICD-9 and ICD-10 code. Where the new version is capable to classify diseases more precisely than the old. Figure 5.11 depicted that in ICD-9, we know what portion of the breast involved that is central portion, but we do not know in which breast. In ICD-10, we not only know which portion of the breast is involved, but we also know whether it is the right or left breast.

¹⁸<https://www.cms.gov>

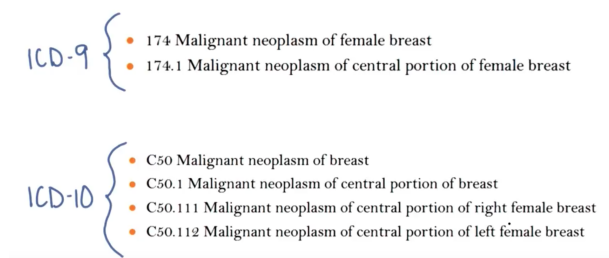


Figure 5.11: Comparison of ICD-9 and ICD-10

Chapter 6

Evaluation

“Truth can be stated in a thousand different ways, yet each one can be true.”

—*Swami Vivekananda*

We conducted a user study to assess the usability of methodology for domain modeling. During an entire period of a semester, students were asked to perform a series steps described in the Chapter 2 to design a domain ontology as an assignment for Knowledge and Data Integration Course (KDI¹). At the end of the semester, we asked all the participants to fill an online questionnaire. An open-ended group discussion was also conducted at the end of the session. Data from the discussion session were further analyzed and mapped with the online questionnaire. The questionnaire was designed to understand different UX dimensions [Laugwitz et al., 2008] along with the specific traits of the methodology.

[Laugwitz et al., 2008] came up with more generic User Experience dimensions. They are:

1. ***Attractiveness*** relates to the overall impression of the system

¹<http://disi.unitn.it/ldkr/>

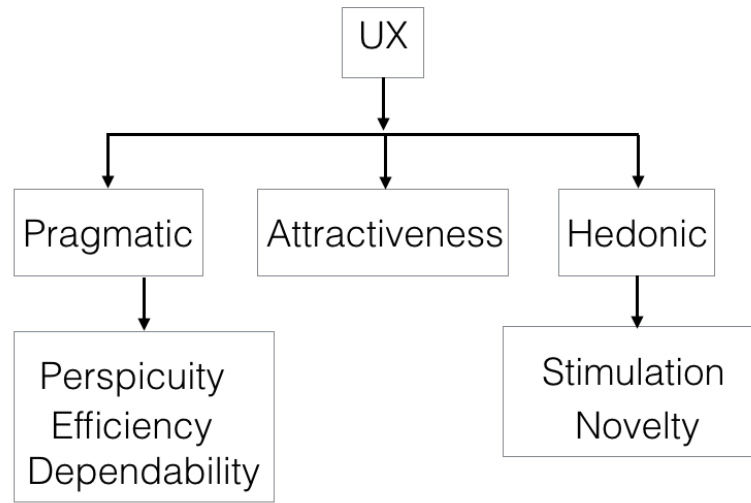


Figure 6.1: User Experience dimensions. Adapted from [Laugwitz et al., 2008].

2. ***Pragmatic*** It measures the usefulness of the system across three dimensions. Efficiency deals with how easily the system can be used. Perspicuity deals with the familiarity of the system. Finally dependability deals with the user's feeling of the control of the system.
3. ***Hedonic*** Users stimulation is measured in this dimension. It is composed of stimulation (deals with feeling of excitement while using the system)and Novelty (The innovativeness of the system.).

These UX dimensions perform a thorough assessment of the product using five scales with 20 terms. These scales were: Perspicuity, Efficiency, Dependability, Stimulation, and Novelty. These scales were further categorized as Pragmatic (the usability of the system and basically consists of Efficiency, Perspicuity, and Dependability) and Hedonic (user stimulation while seeing the product) quality. It consists of 2 scales: Stimulation and Novelty. After that, to validate the usefulness of the methodology, we asked the participants to point out the advantages and disadvantages of the the methodology.

A total number of 18 participants (13 Male, 5 Female; 14 within an age range of 18-25 and 4 within an age range of 26-30) took part in the user study. Participants' highest degree of education ranged from undergraduate (3) to postgraduate (15) degree. 17 of them were studying computer science whereas one was from linguistics department. Some of the participants were fully aware of the semantic technologies and tools whereas others had no prior knowledge regarding semantics. All of the participants possessed good knowledge of English and volunteered for the study. The consent form was signed prior to the beginning of the study. Throughout the semester the students were asked to model an ontology based on the methodology.

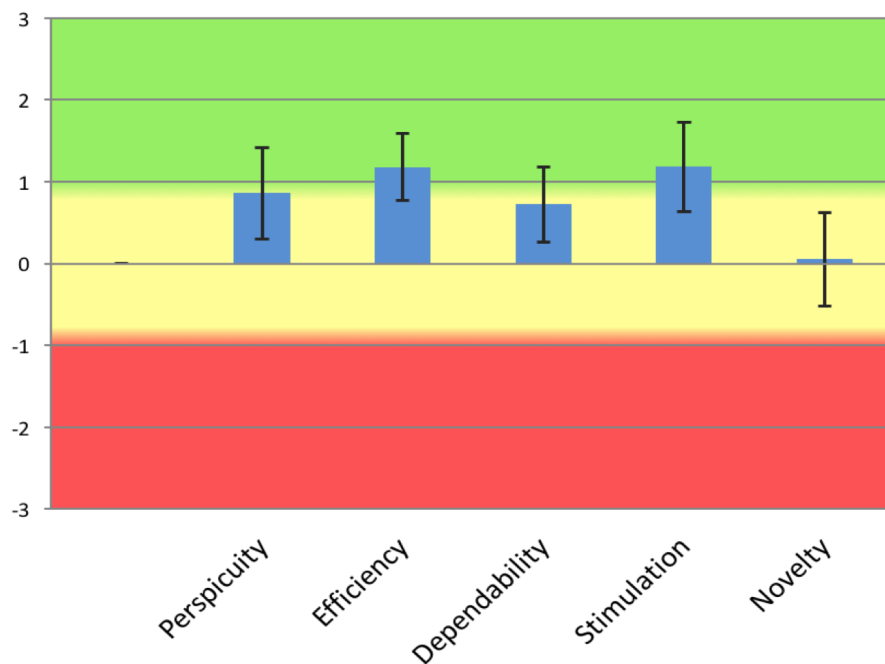


Figure 6.2: Result of Methodology Evaluation

The result of user evaluation² shows (see Figure 6.2) that the age, qualification and location has no impact on using the methodology. Participants' main motivation was to get good marks on the course and also

²<https://goo.gl/5BVK3u>

to learn data integration. The result obtained from the participants were mostly positive with a mean value of (0.86) for perspicuity, (1.181) for efficiency, (0.722) dependability, (1.185) stimulation and (0.056) for novelty. The overall pragmatic quality was 0.92 where as the hedonic quality was 0.63. This shows that, the methodology was user friendly and easy to learn. However, the participant didn't find it stimulating and novel. The elaborated examples and one one to one teacher- student interaction made the methodology easy. The practical benefit were also noticed by the participants. The participants also thought there were some issues, some participants were unable to understand the benefit of designing ontologies. The formalization process of DERA into Description Logic was also considered difficult. Similarly, some participants felt that not all the generalized queries were answered by the model. Some participants felt more emphasis has to be made on the linguistic level rather than generalized queries. Some felt development life cycle can be more streamlined by using the results.

Considering the suggestions from the participants, we will plan to perform a new comparative evaluation. Where similar questionnaire will be asked based on five UX dimensions to a new set of students taking the course. After that we will perform the comparative analysis over the data and test the efficiency of the methodology.

Chapter 7

Implementation

“To develop a complete mind:
Study the science of art; Study
the art of science. Learn how to
see. Realize that everything
connects to everything else.”
—*Leonardo da Vinci*

The best way to verify a model or a theory is through implementation. As Fernández-López said “*Obviously, if ontologies are to be used by computer, they have to be implemented*” [Fernández-López, 1999].

The multi-layered approach is adapted from the software engineering principle which allows easy maintainability and encapsulation of data (entities) that will help in creation of a high performance, generic and adaptive systems. The architecture is presented in the Figure 7.1. The layers are:

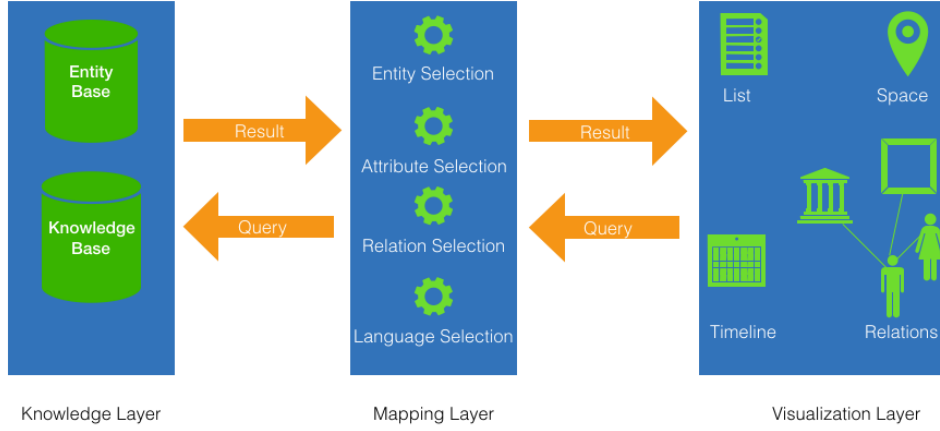


Figure 7.1: Layered Architecture

- ***The Knowledge Layer*** It stores data as entities which are interconnected with each other forming a dense entity graph. In this graph, the nodes represent the entities and the links represent the relations between them. The data are modeled as entities along with their attribute and relations following an entity-centric approach using an eType.
- ***The Visualization Layer*** It visualizes entities from various contextual perspectives such as timeline, space, network and list. Furthermore, the User Interface allows visualization of an entity 1) as a whole with all of its attributes and relations and 2) also according to the users' own representation of the entity. This adaptivity contributes towards developing an intuitive, natural and generic User Interface for a vast number of users.
- ***The Logical Mapping Layer-*** This layer is crucial because an entity graph is dense and the user will be overwhelmed with an information overload on the User Interface. Therefore, it interacts with both *Knowledge layer* and *Visualization layer* and fetches specific parts of

the entity graph from the *knowledge layer* according to the request issued from the *visualization layer*.

We implemented our proposed framework by developing a user friendly visualization tool named *SemUI* (Semantic User Interface). It is a semantic multiview visualization tool. The development process of the tool is not the scope of this work. SemUI (see Figure 7.2) has multiple views as a timeline (c), space (d), network (e) and tabular (f) views. Search box (a) allows direct entity search. eType hierarchy facilitates exploration (b). History option facilitates traversing back to the previously explored entity or eType (g).

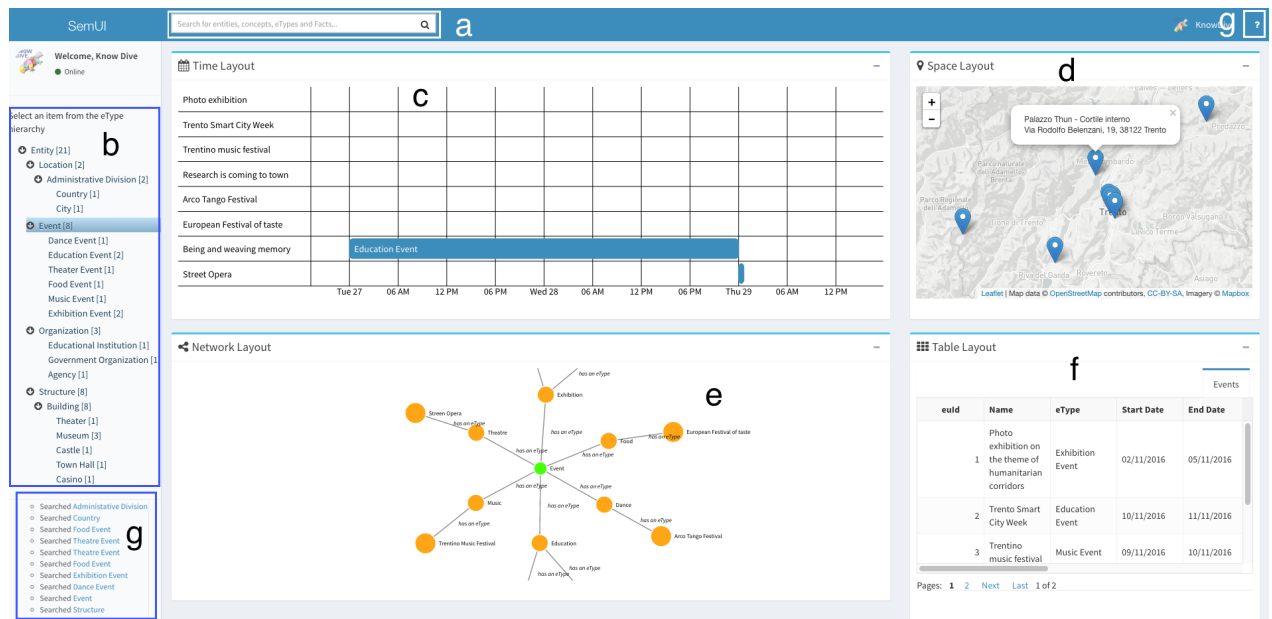


Figure 7.2: SemUI

Different section of the UI accommodates different components of our proposed multi-layered framework. For instance, left panel of the UI (see Figure 7.2 b) shows how an eType is ontologically related with its super-classes and subclasses. This view allows user to understand the position of their chosen eType in the hierarchy. Upper right section (see Figure 7.2d)

of the panel accommodate all spatial attribute and represent on the map view. Below the map view, user can see a list of all entity belong to the chosen eType (In our case all events) in a table view (see Figure 7.2f). Top-middle section of the UI cater for temporal attribute and present in a timeline view (see Figure 7.2c). The Network view which display inter connected entities and their relation (see Figure 7.2e). This is very pleasant view of the UI and most of the user like this way of representation (for details result see the link¹).

The second implementation on healthcare domain implemented in the *SHIB-UI*².

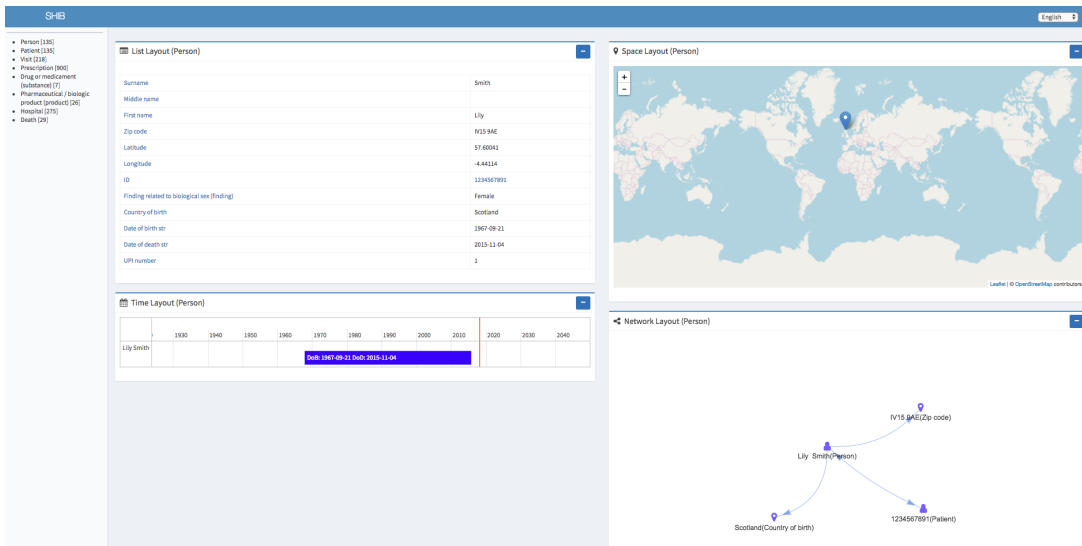


Figure 7.3: SHIB User Interface

¹<https://goo.gl/bpNASM>

²<http://opendata-staging.disi.unitn.it:54185/patient/id/1234567891/language/en>

Chapter 8

Conclusion and Future Work

“I do the very best I know how -
the very best I can; and I mean
to keep on doing so until the
end.” —*Abraham Lincoln*

In this thesis, in order to harmonize diversity problems, we proposed a common data model for defining and managing diverse information of domain-specific data. Firstly, we described various components that a domain should have and then we introduced a schema and proposed a formal and generic ontology-based model to capture the diversity. Secondly, we defined a vocabulary which links between schema and terms used in the model. The proposed framework improves integration process by accommodating different schemas. Our architecture has been successfully implemented in a real system called *SemUI*¹ (see Figure 7.2). The methodology has been validated on three different domain applications as well as some within the same domain and the results obtained are promising.

In the geospatial domain

Healthcare domain is a very mature domain.

¹<http://opendata-demo.disi.unitn.it:34500/ep/>

8.1 observation

Before any technology development begins, we must first determine the required components through an interactive design process. Here we have not provided a detailed process manual to be followed step by step, nor is it an explicit implementation of the various “design thinking, “human centered design, or “integrative thinking processes, which are excellent models to follow. Instead, we present here the crucial recommendations suggested by the Health informatics course run by MITx², which we need to consider when integrating an informatics solution in a global health environment as well as relevant for other intercontinental integration task. The primary focus is on quality improvement brings attention to capturing and communicating the right information, while also connecting to the workflow loop for appropriate clinical intervention. Equally important, we must focus on making applications easy to use in order to encourage adoption and sustainability. In order to accomplish these challenging, often conflicting goals, we recommend incorporating the following mechanisms into your design process.

8.1.1 Stakeholder Analysis

Understanding the key players and users involved with an intervention is central to deploying a usable and effective solution. This goes beyond just end users to include anyone that is involved in system the delivery process for your intervention.

When considering other stakeholders to design for, it is important to be comprehensive in considering who else is in the process. This means not just individual end users of the technology but any individual or organization involved in the process, directly or indirectly, since they can

²<https://courses.edx.org/courses/course-v1:MITx+HST.936x+1T2017/course/>

have a significant influence on the design of the solution. It is helpful to consider any group that can influence the funding, implementation, or public policy around an intervention. Common stakeholders to recognize include providers who deliver care, payers who reimburse for procedures or care delivery, funding agencies who may fund the overall pilot program, or governmental organizations that establish public policy or drive health initiatives. Each of these groups can have a significant impact on the sustainable success of an intervention, while not considering any single one can obstruct project progress regardless of the success of other aspects. For example, even with a successful clinical pilot and evidence of quality improvement, if the local government does not provide supportive public policies or instantiates prohibitive policies in regard to data or treatment regulations, an intervention cannot be scaled up.

8.1.2 Workflow Integration

Any technical innovation needs to be thoughtfully integrated within the workflow. Rather than designing around a novel technology, its often constructive to begin with the clinical intervention and analyze where technology can play a role and improve processes. At the outset, it is helpful to investigate any significant barriers to implementation, such as preexisting systems, user hesitance, immutable processes, or insufficient human resources.

An aspect of workflow integration is a strong emphasis on workflow efficiency. Clinicians are notoriously overworked, with limited time and significant amounts spent on documentation already. As much as possible, any innovation needs to avoid adding work or the solution wont be adopted by the clinician user base, nearly regardless of the potential benefit. Therefore, it is imperative to design efficient clinical interfaces that minimize the amount of additional work, with improving efficiency an ideal objective.

An important caveat comes from including the previous recommendation of end-user analysis, where the clinician users will likely have different biases toward what interface improves efficiency from the implementing engineers, so it is prudent to include clinicians in the design process.

Most importantly, for any mHealth solution to yield a lasting impact, it must be directly tied to quality improvement. This requires closing the loop with a clinical intervention, so it is strongly recommended that a holistically designed intervention connect the informatics innovation with the clinical process. This is a challenging proposition, and it requires clinicians to collaborate closely with technologists to push the envelope of combining what is technically feasible with what can make a clinical impact.

8.2 Lesson Learned

Our main goal to enable interaction between domain-specific resources with general purpose resources and to support semantic interoperability in applications, we need to enriched WordNet with a huge number of domain-specific concepts. We give particular attention not only to the quantity, but also to the quality of the information being integrated. Towards this goal we propose set of guiding principles to develop the domains which we learned while working with different domains.

8.2.1 Extending the WordNet with Domain Knowledge

A new domain-specific resource (DSR) initially consists of the top-level Common Sense Knowledge (CSK) i.e. WordNet. On level-2, this corresponds to a mirror of the WordNet synset graph, and is expressed in a set of natural languages in level-1. The DSR is then extended by language and domain resources, the former imported into level-1 and the latter into level-2, attached to the existing DSR concept subtree through

concept equivalence and hypernymy. Typically, domain resources provide both conceptual (level-2) and linguistic (level-1) representations. Level 2 is always imported first, and then the corresponding level 1 representations in one or more languages.

8.2.2 Deciding the Import Order

To each domain-specific resource (DSR), a priority is associated such as that a total order among priorities can be established. As a first approach, the priority is defined as the authority of the resource. However, other factors may also be possible to take into account, such as the size of the resource or its generality or specificity.

8.2.3 Attaching Domain sub-tree

Two sub-trees, which may be the CSK and a domain sub-tree, or two domain sub-tree, depending on the order in which resources are processed are attached together based on equivalence or subsumption relations holding between concepts on either side. While the preferred type of evidence for attaching domain resources is concept equivalence, there are at least two common cases where equivalences could not be available: (i) if the domain resource is highly specialized, the General Knowledge Core (GKC) may not contain an equivalent even to its most general (i.e., root) concept; (ii) if the domain resource contains concepts that are either complex concepts or correspond to a different modelling of common-sense knowledge than what is provided by the GKC.

8.2.4 Specifying Concept Metadata

We need to specify the part of speech of the importing term. It is usually (but not always) noun for domain terms;

Provenance (also called reference in the importer) information needed to be indicated both for concepts and for concept relations, and should refer to the domain resource; Optional attributes: description (glossary): this should in principle provided by the domain resource; in its absence, the field can be left blank; concept label: a single concise English word or expression describing the concept, this can be generated from the rank 1 lemma of the corresponding English synset; For example, in case of concept BMI, which has three lemma *Body mass index*, *BMI* and *Z68*, concept label is *Body mass index*. note: this is to be indicated both for concepts and for concept relations.

8.3 Future Work

In our future study, we will consider to expand our methodology as well as eType model to design more domain applications. Many different domain adaptations, tests, and experiments have been left out for the future due to the time constrain. Future work concerns deeper analysis of various domains, maintain various versions and proposals to adapt different methods.

Bibliography

J Akoka, L Berti-Equille, O Boucelma, Mokrane Bouzeghoub, I Comyn-Wattiau, M Cosquer, V Goasdoué-Thion, Z Kedad, S Nugier, V Peralta, et al. A framework for quality evaluation in data integration systems. In *9th International Conference on Enterprise Information Systems (ICEIS)*, page 10, 2007.

Robert Arp and Barry Smith. Function, role, and disposition in basic formal ontology. In *The 11th Annual Bio-Ontologies Meeting*, page 1, 2008.

American Medical Association. *Current procedural terminology: CPT*. American Medical Association, 2007.

British Medical Association et al. British national formulary: 38. 1999.

European Language Resources Association. What is a language resource?, 2017. URL <http://www.elra.info/en/about/what-language-resource/>.

Franz Baader. *The description logic handbook: theory, implementation, and applications*. Cambridge university press, 2003.

Miriam Baglioni, Maria Vittoria Masserotti, Chiara Renso, and Laura Spinsanti. Improving geodatabase semantic querying exploiting ontologies. In *GeoSpatial Semantics*, pages 16–33. Springer, 2011.

Perakath C Benjamin, Christopher P Menzel, Richard J Mayer, Florence Fillion, Michael T Futrell, Paula S deWitte, and Madhavi Lingineni. Idef5 method report. *Knowledge Based Systems, Inc*, 1994.

Tim Benson. Principles of health interoperability hl7 and snomed; 2009.

Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. Revising the wordnet domains hierarchy: semantics, coverage and balancing. In *Proceedings of the Workshop on Multilingual Linguistic Resources*, pages 101–108. Association for Computational Linguistics, 2004.

Ronald J Brachman, Deborah L McGuinness, Peter F Patel-Schneider, Lori Alperin Resnick, and Alexander Borgida. Living with classic: When and how to use a kl-one-like language. *Principles of semantic networks*, pages 401–456, 1991.

D Brickley. Wgs84 geo positioning: an rdf vocabulary, 2004.

Elliot G Brown, Louise Wood, and Sue Wood. The medical dictionary for regulatory activities (meddra). *Drug safety*, 20(2):109–117, 1999.

Ancient World Mapping Center, the Stoa Consortium, and the Institute for the Study of the Ancient World. Pleiades, 2014. URL <http://pleiades.stoa.org/home>.

Uppsala Monitoring Centre. The who adverse drug reaction terminology, 2010. URL <https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/WHO/>.

Sandeep Chaware and Srikantha Rao. Integrated approach to ontology development methodology with case study. *International Journal of Database Management Systems*, 2(3):13–19, 2010.

- Peter Pin-Shan Chen. The entity-relationship model toward a unified view of data. *ACM Transactions on Database Systems (TODS)*, 1(1):9–36, 1976.
- Mihai Codescu, Gregor Horsinka, Oliver Kutz, Till Mossakowski, and Rafaela Rau. Do-roam: activity-oriented search and navigation with openstreetmap. In *GeoSpatial Semantics*, pages 88–107. Springer, 2011.
- European Commission. European Interoperability Framework (EIF) for european public services, 2010. URL http://ec.europa.eu/isa/documents/isa_annex_ii_eif_en.pdf.
- Federal Geographic Data Committee. Entity and attribute information, 2015. URL <https://www.fgdc.gov/metadata/csdgm-new/05.html>.
- INSPIRE Directive. Directive 2007/2/ec of the european parliament and of the council of 14 march 2007 establishing an infrastructure for spatial information in the european community (inspire). *Published in the official Journal on the 25th April*, 2007.
- Biswanath Dutta, Usashi Chatterjee, and Devika P Madalli. Yamo: Yet another methodology for large-scale faceted ontology construction. *Journal of Knowledge Management*, 19(1):6–24, 2015.
- Thomas J Eggebraaten, Jeffrey W Tenner, and Joel C Dubbels. A health-care data model based on the hl7 reference information model. *IBM Systems Journal*, 46(1):5–18, 2007.
- Mariano Fernández-López. Overview of methodologies for building ontologies. 1999.
- Mariano Fernández-López, Asunción Gómez-Pérez, and Natalia Juristo. *Methontology: from ontological art towards ontological engineering*. American Association for Artificial Intelligence, 1997.

Food and Drug Administration (FDA). National drug code directory. <https://www.fda.gov/Drugs/InformationOnDrugs/ucm142438.htm>, 2017.

April Fritz, Constance Percy, Andrew Jack, Kanagaratnam Shanmugaratnam, Leslie Sobin, D Max Parkin, Sharon Whelan, et al. *International classification of diseases for oncology*. Number Ed. 3. World Health Organization, 2000.

Patricia J. Gagnon and Martha Sarbey De Souto. *Travel career development*. The travel institute, 2016.

Fausto Giunchiglia and Biswanath Dutta. Dera: A faceted knowledge organization framework. Technical report, University of Trento, 2011.

Fausto GIUNCHIGLIA and Mattia FUMAGALLI. Concepts as (recognition) abilities¹. In *Formal Ontology in Information Systems: Proceedings of the 9th International Conference (FOIS 2016)*, volume 283, page 153. IOS Press, 2016.

Fausto Giunchiglia, Vincenzo Maltese, Feroz Farazi, and Biswanath Dutta. Geowordnet: a resource for geo-spatial applications. In *The Semantic Web: Research and Applications*, pages 121–136. Springer, 2010.

Fausto Giunchiglia, Biswanath Dutta, and Vincenzo Maltese. From knowledge organization to knowledge representation. *Knowledge Organization*, 41(1), 2014.

Asunción Gómez-Pérez. Evaluation of ontologies. *International Journal of intelligent systems*, 16(3):391–409, 2001.

Asunción Gómez-Pérez, Natalia Juristo, and Juan Pazos. Evaluation and assessment of knowledge sharing technology. *Towards very large knowledge bases*, pages 289–296, 1995.

Google. Gtfs static overview, 2006.

Thomas R Gruber. Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies*, 43(5-6):907–928, 1995.

M Gruninger. Methodology for the design and evaluation of ontologies. In *Proc. IJCAI’95, Workshop on Basic Ontological Issues in Knowledge Sharing*, 1995.

Nicola Guarino. *Formal ontology in information systems: Proceedings of the first international conference (FOIS’98), June 6-8, Trento, Italy*, volume 46. IOS press, 1998.

Nicola Guarino and Christopher Welty. Evaluating ontological decisions with ontoclean. *Communications of the ACM*, 45(2):61–65, 2002.

Giancarlo Guizzardi, Ricardo de Almeida Falbo, and Renata SS Guizzardi. Grounding software domain ontologies in the unified foundational ontology (ufo): The case of the ode software process ontology. In *CIbSE*, pages 127–140, 2008.

Mordechai Haklay and Patrick Weber. Openstreetmap: User-generated street maps. *Pervasive Computing, IEEE*, 7(4):12–18, 2008.

Crispin Hales. *Analysis of the engineering design process in an industrial context*. PhD thesis, University of Cambridge, 1987.

Gergely Héja, György Surján, and Péter Varga. Ontological analysis of snomed ct. *BMC medical informatics and decision making*, 8(1):S8, 2008.

Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, and Gerhard Weikum. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 3161–3165. AAAI Press, 2013.

- Marta Iglesias-Sucasas, Soonho Kim, and Virginie Viollier. The fao geopolitical ontology: a reference for country-based information. *Food and Agriculture Organization of the United Nations*, 2013.
- SNOMED International. Snomed international snomed ct browser, 2017. URL <http://browser.ihtsdotools.org/>
- Faten Kharbat and Haya El-Ghalayini. *Building Ontology from Knowledge Base Systems*. INTECH Open Access Publisher, 2008.
- Kevin Knight, Ishwar Chander, Matthew Haines, Vasileios Hatzivasiloglou, Eduard Hovy, Masayo Iida, Steve K Luk, Richard Whitney, and Kenji Yamada. Filling knowledge gaps in a broad-coverage machine translation system. *arXiv preprint cmp-lg/9506009*, 1995.
- Dave Kolas and R Batle. Geosparql user guide, 2012.
- Bettina Laugwitz, Theo Held, and Martin Schrepp. Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and Usability Engineering Group*, pages 63–76. Springer, 2008.
- Vivian Law, Craig Knox, Yannick Djoumbou, Tim Jewison, An Chi Guo, Yifeng Liu, Adam Maciejewski, David Arndt, Michael Wilson, Vanessa Neveu, et al. Drugbank 4.0: shedding new light on drug metabolism. *Nucleic acids research*, 42(D1):D1091–D1097, 2014.
- Jon Jay Le Grange, Jens Lehmann, Spiros Athanasiou, A Garcia-Rojas, Giorgos Giannopoulos, Daniel Hladky, Robert Isele, A-C Ngonga Ngomo, M Ahmed Sherif, Claus Stadler, et al. The geoknow generator: Managing geospatial data in the linked data web. *Linking Geospatial Data*, 2014.
- Simon Liu, Wei Ma, Robin Moore, Vikraman Ganesan, and Stuart Nelson. Rxnorm: prescription for electronic drug information exchange. *IT professional*, 7(5):17–23, 2005.

- Yiqi Liu and Masaharu Yoshioka. Construction of large geographical database by merging wikipedias geo-entities and geonames. Technical report, Tech Report SIG-SWO-A1102-03, 2011.
- Franco Maggio. The italian cadastre and the real estate taxation, 2012. URL https://nexa.polito.it/nexacenterfiles/italian_cadastre.pdf.
- Farzaneh Mahdisoltani, Joanna Biega, and Fabian Suchanek. Yago3: A knowledge base from multilingual wikipedias. In *7th Biennial Conference on Innovative Data Systems Research*. CIDR 2015, 2014.
- George Miller and Christiane Fellbaum. Wordnet: An electronic lexical database, 1998.
- United Nation. Future trends in geospatial information management: the five to ten year vision, 2013. URL <http://ggim.un.org/docs/Future-trends.pdf>.
- United Nation. Unwto tourism highlights, 2016. URL <https://www.e-unwto.org/doi/pdf/10.18111/9789284418145>.
- Natalya F Noy, Deborah L McGuinness, et al. Ontology development 101: A guide to creating your first ontology, 2001.
- Natasha Noy, Alan Rector, Pat Hayes, and Chris Welty. Defining n-ary relations on the semantic web. *W3C Working Group Note*, 12:4, 2006.
- Royal College of General Practitioners et al. Office of population censuses and surveys, department of health. *Morbidity statistics from general practice. Fourth national study*, 1992, 1991.
- Annika Öhgren and Kurt Sandkuhl. Towards a methodology for ontology development in small and medium-sized enterprises. In *IADIS AC*, pages 369–376, 2005.

- World Health Organization. International Classification of Diseases for Oncology , 2013. URL <http://codes.iarc.fr/home>.
- World Health Organization. International statistical classification of diseases and related health problems, 2017. URL <http://www.icd10data.com/ICD10CM/Codes>.
- Matthew Perry and John Herring. Ogc geosparql-a geographic query language for rdf data. *OGC Implementation Standard*, ref: OGC, 2012.
- Oxford University Press. Oxford dictionaries online (odo), 2015. URL <http://www.oxforddictionaries.com>.
- James Pustejovsky, Jessica L Moszkowicz, and Marc Verhagen. Iso-space: The annotation of spatial information in language. In *Proceedings of the Sixth Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 1–9, 2011.
- Shiyali Ramamrita Ranganathan. Prolegomena to library classification. *The Five Laws of Library Science*, 1967.
- Christian Reich, Patrick Ryan, Rimma Belenkaya, Karthik Natarajan, and Clair Blacketer. Omop common data model v5.1 specifications, 2017. URL <https://github.com/OHDSI/CommonDataModel/wiki>.
- James C Reichert, Matt Glasgow, Scott P Narus, and Paul D Clayton. Using loinc to link an emr to the pertinent paragraph in a structured reference knowledge base. In *Proceedings of the AMIA Symposium*, page 652. American Medical Informatics Association, 2002.
- Philip Sargent, Eswaran Subrahmanian, Mary Downs, Reid Greene, and Diane Rishel. Materials information and conceptual data modeling. In *Computerization and Networking of Materials Databases: Third Volume*. ASTM International, 1992.

- Barry Smith and Achille C Varzi. Fiat and bona fide boundaries. *Philosophical and Phenomenological Research*, pages 401–420, 2000.
- ASTM Standard. E2369-05e1 standard specification for continuity of care record (ccr). *West Conshohocken, PA: ASTM International*, 2005.
- EUROPEAN COMMITTEE FOR STANDARDIZATION. Public transport - Service interface for real-time information relating to public transport operations - Part 4: Functional service interfaces: Facility monitoring, 2011. URL http://www.kti.hu/uploads/KMK/2011/SIRI%20Tudstr/Szabvny%20UK/SIRI_15531-4.pdf.
- Eswaran Subrahmanian, Suresh L Konda, Sean N Levy, Yoram Reich, Arthur W Westerberg, and Ira Monarch. Equations arent enough: Informal modeling in design. *AI EDAM*, 7(4):257–274, 1993.
- Bill Swartout, Ramesh Patil, Kevin Knight, and Tom Russ. Toward distributed use of large-scale ontologies. In *Proc. of the Tenth Workshop on Knowledge Acquisition for Knowledge-Based Systems*, pages 138–148, 1996.
- Yafang Wang, Mingjie Zhu, Lizhen Qu, Marc Spaniol, and Gerhard Weikum. Timely yago: harvesting, querying, and visualizing temporal knowledge from wikipedia. In *Proceedings of the 13th International Conference on Extending Database Technology*, pages 697–700. ACM, 2010.
- Margriet Nieuwenhuis Eva van Steenbergen Wendy Carrara, Wander Engbers. Analytical Report 4: Open data in cities, 2016. URL http://www.europeandataportal.eu/sites/default/files/edp_analytical_report_n4_-_open_data_in_cities_v1.0_final.pdf.

Mark Wick and Bernard Vatant. The geonames geographical database.
Available from World Wide Web: <http://geonames.org>, 2012.

Appendix A

Authority File

“I feel like I a lot to express, and
I'm not gifted.” —*Cristina,*
Vicky Cristina Barcelona

A.1 Tools

Here we listed all tools that are used during our work. They are:

Protégé: Protégé¹ is a free, open source ontology editor and a knowledge management system. Protégé provides a graphic user interface to define ontologies. It is use for creating model, consistency checking, importing and integrating others OWL ontologies.

HermiT: HermiT² is reasoner for ontologies written using the Web Ontology Language (OWL). Given an OWL file, HermiT can determine whether or not the ontology is consistent, identify subsumption relationships between classes, and much more. In our work, we mainly use HermiT reasoner to check the consistency of the model.

GraphDB: GraphDB³ is an enterprise ready Semantic Graph Database,

¹<https://protege.stanford.edu>

²<http://www.hermit-reasoner.com>

³<http://graphdb.ontotext.com>

compliant with W3C Standards. Semantic graph databases (also called RDF triplestores) provide the core infrastructure for solutions where modeling agility, data integration, relationship exploration and cross-enterprise data publishing and consumption are important. GraphDB has been used to perform SPARQL query, analytics, knowledge graph and class relationship.

yEd: yEd⁴ is a free general-purpose diagramming program with a multi-document interface. It is a cross-platform application written in Java that runs on Windows, Linux, Mac OS, and other platforms that support the Java Virtual Machine. yEd has been used to draw all ExER diagrams.

A.2 Geo Authority File

The space domain is a very important, where the various stakeholders such as the general public and policy makers require authentic data to run their day-to-day business or to mitigate the emergency situation. Furthermore, cross border information (geographical features which are shared by multiple administrative units such as river, mountain) requires highly reliable data to manage different application and policy-making decisions. So, it becomes vital to collect authentic data from authoritative sources. Here we enlisted such reliable, authoritative sources.

⁴<https://www.yworks.com/products/yed>

Table A.1: OpenStreetMap

Title	Open Street Map
Acronym	OSM
URL/Website	https://www.openstreetmap.org
Description	OSM represents physical features on the earths surface using tags.
eTypes	Space region (Administrative unit, Transportation link, Transportation point, Body of water, protected site) Artifact (Building, Structure) Facility
Attribute	Geographical name, landform, vegetation, Longitude, latitude, address, contact, land use, elevation etc.
Size	1.92 billion
Coverage	Global
Format	Geojson, SHP, PBF, XML
Frequency of update	Weekly
URL	https://www.openstreetmap.org

A.3 eTypes Specification

All Geo eTypes and their respective attributes can be found in a excel format in the link <https://goo.gl/BnDqtx>.

Other Healthcare eTypes and their attributes are shown from Table A.3 to A.7

Table A.2: GeoNames

Title	GeoNames
Acronym	OSM
URL/Website	https://www.openstreetmap.org
Description	OSM represents physical features on the earths surface using tags.
eTypes	Space region (Administrative unit, Transportation link, Transportation point, Body of water, protected site) Artifact (Building, Structure) Facility
Attribute	Geographical name, landform, vegetation, Longitude, latitude, address, contact, land use, elevation etc.
Size	10 Million
Coverage	Global
Format	Geojson, SHP, PBF, XML
Frequency of update	Weekly
URL	https://www.openstreetmap.org

Table A.3: Specimen eType

AttributeName	DataType	Standard
specimen of person	eType (Person)	OMOP
specimen concept	concept	OMOP
specimen collection date	Date	OMOP SNOMED-CT
specimen type	concept	OMOP
quantity	Float	OMOP
unit	concept	OMOP
anatomic site	eType (body structure)	OMOP
disease status	concept	OMOP

Table A.4: Death eType

AttributeName	DataType	Standard
death of person	Etype (Person)	OMOP
date of death	date	OMOP SNOMED-CT
death datetime	datetime	OMOP
cause of death	concept	OMOP

Table A.5: Visit eType

AttributeName	DataType	Standard
person visit	eType (Person)	OMOP
visit start date	date	OMOP SNOMED-CT
visit start datetime	datetime	OMOP
visit end date	date	OMOP
visit end datetime	datetime	OMOP
care site visited	eType (healthcare facility)	OMOP
healthcare provider	eType (Health care personnel)	OMOP

Table A.6: Provider eType

AttributeName	DataType	Standard
Drug Enforcement Administration Number	string	OMOP
specialty	concept	OMOP
practice site	eType (care site)	OMOP INSPIRE
National Provider Identifier	string	OMOP
role of	eType (Person)	OMOP

Table A.7: Disease eType

AttributeName	DataType	Standard
finding site	eType (body structure)	SNOMED-CT
associated morphology	eType (body structure)	SNOMED-CT
after	eType (procedure)	SNOMED-CT
causative agent	eType (organism or substance)	SNOMED-CT
due to	eType (disease or procedure)	SNOMED-CT
severity	concept	SNOMED-CT
occurrence	concept	SNOMED-CT
has definitional manifestation	concept	SNOMED-CT
finding method	eType (procedure)	SNOMED-CT
finding informer	eType (person)	SNOMED-CT